26-28

April 2007-Jan. 2008

विश्वभारत @ tdil

VishwaBharat

TDIL PROGRAMME
VISION
to
MISSION

Language Technology Flash

VishwaBharat@tdil

April 2007 - January 2008

Patron

Shri Jainder Singh, Secretary
Department of Information Technology
secretary@mit.gov.in

Advisor

Shri E.K. Bharat Bhushan
Joint Secretary and Financial Advisor,
Department of Information Technology
bharatbhushan@mit.gov.in

Editorial....

MESSAGE

The year 2005 witnessed a major initiative of the Government linking people through Indian Language Technologies by releasing software tools and fonts for Indian Languages such as Hindi, Tamil and Telugu followed by release of 7 more language CDs for Assamese, Kannada, Malayalam, Marathi, Oriya, Punjabi and Urdu in January, 2007 and with the commitment to release such CDs in the balance officially recognized Languages by March, 2008 through National Roll Out Plan Project being implemented at C-DAC, Pune. In view of the social impact of the IT Technology on the life of the people and to bridge the digital divide being created, the Govt. of India enabled all the citizens of the country to receive the benefits of the Information Technology revolution through making the CDs available for free use by downloading from the websites http://www.ildc.in & http://www.ildc.gov.in and also getting the CD on request through these websites.

The CDs contain Basic Information Processing Kit (BIPKs) which includes localized Bharteeya Open Office, Fire Fox Browser and E-mail Client, Open Type Fonts, Keyboard drivers etc. The CDs also include tools like Spell Checker, Dictionary, Code Converters, Typing Assistant , Language Learning Tool, etc. & some of the lab scale technologies developed under TDIL Programme such as Machine Translation systems, OCRs etc. for feed back and improvisation

This has been achieved through contributions of research efforts undertaken through TDIL Programme and various contributions by the Public and Private players, academicians, R&D and Industry.

This issue details out the contents of the Language CDs for which a large user base has already been created in the country and these have been very well received by the social community. This initiative will go a long way in bridging the digital divide which to a larger extent has also been created by the language technology barrier.

(E.K. Bharat Bhushan)

Contents

Editorial Team

ISSN No. 0972-6454 has been granted to VishwaBharat@tdil. Google search engine refers to the contents of this journal
Website : www.tdil.mit.gov.in
Cover Design Shri Narendra Shrivastava

1. Calendar of Events

1. IJCNLP 2008, The Third International Joint Conference on Natural Language Processing. Hyderabad, India, **January 7-12, 2008.** http://www.ijcnlp2008.org/

2. IJCNLP 2008 The 6th Workshop on Asian Language Resources. Hyderabad, India, **January 11-12, 2008.** http://tanaka-www.cs.titech.ac.jp/ALR/WS/6th/accepted.htm

3. INFOS2008 The 6th International Conference on Informatics and Systems. Faculty of Computers and Information. Cairo University 5 Dr. Ahmed Zoweil st., Dokki, Giza 12613, Egypt. **March 27-28 2008.** http://www.fci.cu.edu.eg/INFOS2008/

4. ICASSP 2008 - The 33rd International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, United States. **30 March 2008 - 04 Apr 2008** http://www.icassp2008.org/CallForPapers.asp

5. 5th International Conference on Information Technology : New Generations ITNG 2008 Las Vegas, Nevada, USA **April 7-9, 2008.** www.itng.info

6. ANLP at FLAIRS 2008 — Applied Natural Language Processing. Miami, United States. **May 15-17, 2008.** http://www.msstate.edu/dept/ english/applied_nlp/flairs_2008

7. The 5th European Semantic Web Conference (ESWC 2008) is being held in Tenerife, Spain from **June 1-5, 2008.** http://www.semanticfocus.com/blog/entry/title/eswc-2008-5th-european-semantic-web-conference/

8. The 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008) Barcelona - Spain. **June 12-13, 2008** Source http://www.iceis.org/workshops/nlpcs/ nlpcs2008-cfp.html

9. INLG 2008 - International Natural Language Generation Conference. Salt Fork, Ohio, United States, **June 12-14, 2008.** http://www.ling.ohio-state.edu/inlg2008/

10. ACL demos 2008 - Call for Demos, The 46th Annual meeting of the Association for Computational Linguistics: Human Language Technologies. Columbus, Ohio, United States. **June15 - 20 2008.** http://www.acl2008.org/

11. CIAA 2008 - Thirteenth International Conference on Implementation and Application of Automata (CIAA) San Francisco, United States., **July 21-24, 2008.** http://ciaa2008.cs.sonoma.edu

12. Coling 2008 - The 22nd International Conference on Computational Linguistics. Manchester, UK. **August 18-22, 2008** http://www.informatics.susx. ac.uk/research/groups/nlp/

13. 6th International Conference on Natural Language Processing, GoTAL 2008 Gothenburg, Sweden, **August 25-27, 2008.** http://www.cse.chalmers.se/ gotal.html

14. 2008 ACTFL ANNUAL CONVENTION AND WORLD LANGUAGES EXPO (Pre-convention workshops on Thursday, November 20) Orlando, Florida, **November 21-23, 2008** http://convention2.allacademic.com/one/actfl/actfl08/

Unicode
http://www.unicode.org/timesens/calendar.html

15. UTC # 115 / L2 # 212Hosted by Adobe, San Jose, CA, **May 12 - 16, 2008**

16. UTC # 116 / L2 # 213 Hosted by Microsoft, Redmond, WA, **Augusta 11 - 15, 2008**

17. UTC # 117 / L2 # 214 Unicode Annual Members Meeting Hosted by Google, Mountain View, CA, **November 3 - 7, 2008**

2. Paradigm Shift of Language Technology Initiatives under TDIL Programme

TDIL Programme was launched by the Ministry in the year 1991 by initiating tax corpora development in officially recognized Indian languages. It was well understood that the language technology developed will require coupling of efforts between computer scientists, linguists and language technology experts. Formal academic curriculum is available for such interdisciplinary courses in the country. Therefore natural language process teachers training programmes were conducted across the country to train the computer scientists in the various NLP aspects and also a curriculum was designed for linguists and language teachers to train them on the information technology tools and technologies. For the future teachers of the country, a course was introduced at the B.Ed. level in number of institutions called Computer Aided Learning and Teaching. Many projects were initiated for carrying out fundamental research in NLP especially to explore the use of Sanskrit language which is the mother of all Indian languages.

Thus, the years from 1991 – 2005 went through a number of research projects implementation at various IITs, C-DAC and other academic organizations for carrying out exploratory work in the areas of Machine Translation, Optical Character Recognition, Text to Speech Synthesis, Natural Language Query Systems etc. This lead to the development of number of concept system in these areas. In the year 2000, a need was felt to seed this research activity in all the States and enlarge the geographical spred of researchers working in this area, for which 13 Resource Centres (RCs) were set up across the country with a specific mandate of carrying out research for a particular language and also to build linguistic research tools for all the Indian languages. Therefore, the period form 1991-2000 was the technology initiation phase and we entered to the technology catch up phase in the year 2000. The Resource Centres' activity definitely led to the nucleation of research Centres in this area. Through this project, a large number of institutions were geared to undertake the language technology research. However, it was envisaged that research outputs need to be converted into deployable products. At the Ministerial level in the year 2005, a national level Committee was set up to identify current status in each of the major areas of language technology in India and the world. The composition of the Committee was as under :

1.	Secretary, DIT	Chairman
2.	Prof. N Balakrishnan, IISc., Bangalore	Member
3.	Prof. S Dhande, Director, IIT-Kanpur	Member
4.	Prof. Ramanan, Anna University KBC Centre, Chennai	Member
5.	Prof. Rajeev Sangal, IIIT, Hyderabad	Member
6.	Shri Hemant Darbari, Programme Coordinator, C-DAC, Pune	Member
7.	Shri R Chandrashekhar, JS, DIT	Member
8.	Shri Manoj Annadurai, Shakti Technologies, Chennai	Member
9.	Ms. Swaran Lata, Director, DIT	Member
10.	Rep. of IFD, DIT	Member
11.	Shri V N Shukla, Director (Spl. Applns.), C-DAC, Noida	Member Convenor

The Terms of the reference of the Committee are as follows:

1. Identify current status of language technology in each of the major areas of technology in India and the World.
2. Suggest major technology development goals to be set for TDIL as mission projects and other longer term project (ranging from one to three years) and how such goals can lead to deployable systems.
3. Suggest strategy to be followed to achieve the technology development goals and its deployment.
4. Identify project that can quickly result in products or services that meet public needs.

The Committee deliberated on the initiatives of the TDIL Programme and the impact it created during these years and drafted a road map for undertaking the intensive effort for development and deployment of language technology in the country. The road map is placed at Annexure I.

Follow up initiatives

1. **National Roll out Plan Project :** To consolidate the technologies so far developed under the TDIL Programme and elsewhere in the country, a National Roll-Out Project was initiated by DIT at C-DAC, Pune to bring out the CDs in all 22 Constitutionally recognized languages by providing Basic Information Processing Kits (BIPKs), productivity tools and some of the concept technologies for free use by the public. Under this initiative, 10 language CDs viz. Tamil, Hindi, Telugu, Assamese, Kannada, Malayalam, Marathi, Oriya, Punjabi and Urdu have already been released. This project has lead to uniting people through Indian Languages (refer article on National Rollout Initiative in this issue).

2. **Initiation of R&D in Consortia mode :**
   Based on the experience of implementation of projects at various Centres, it was felt that due to geographical spread of the researchers working in this area, there was a need to undertake activity of R&D in an integrated manner so that the developed research outcomes can be achieved by putting the experience of researchers in this area. There is a need to convert the research outcome to the products by taking the benefit to the masses which involved a fair amount of engineering efforts involving mapping of requirements and exhaustive testing etc. on all laboratory developed products. The Committee identified the focus areas like Machine Translation, Optical Character Recognition, Cross Lingual Information Access and Speech Technology with a time bound result oriented with clear defined products for mass use. The core focus of the consortia approach is **'From Technology to Applications'.**

   The following criteria were defined for choosing applications for development under consortia mode:

   (i) Reaching the public through innovative Indian languages based IT solutions and applications

   (ii) Focusing key application areas E-governance, education, communication, entertainment, e-commerce

   (iii) Concurrent technology development as Indian language Technology is complex.

   (iv) Methodological issues :

   - Separate language specificity from engines
   - Set up limited competitions
   - Evaluation of end-user systems
   - Evaluation of engines
   - Mechanism to seed new groups
   - Manpower development.

Goals for Consortium mode projects:

Short term (6 months – 1 year) – Fonts, Spell Checkers, Office Suite etc. and Setting up of Data Centre Facility to Host resources and applications developed under TDIL Programme and distribution mechanism for Language CDs.

Medium Term (1 – 2 years) – The server based deployment of applications in the following technology domains:

- Machine Translation System from English to Hindi
- Machine translation System among Indian Languages
- Cross Lingual Information Access
- Optical Character Recognition Systems and
- Text and on-line handwriting recognition system.

Long term (2 years and beyond) :

- Research and technology development for the next generation applications such as speech processing, semantic web technology and human computer interaction.

Technology exploration phase

The Requests for proposals (RFP) were floated to initiate the consortia projects in the above listed areas. The focus was to produce language independent engines in the first phase of the Consortia projects which is right now under implementation called the horizontal deliverables. This engine will be vertically integrated by the respective technology groups and 8-10 language pairs were selected in the first phase.

These consortia projects got initiated in the year 2006 and are currently under implementation.

The details of the consortia projects are given below:

Name of the Project	Objectives
Printed Text OCR	• Development of OCR system for printed Indian scripts. • Scripts: Bengali, Devanagari, Malayalam, Gujarati, Telugu, Tamil , Oriya, Tibetan , Nepali, Gurumukhi
On-line Handwriting Recognition system	• Development of On-line Handwriting Recognition System : • Scripts: Bengali, Devanagari, Tamil , Telugu , Kannada and Malayalm Scripts
Cross-lingual Information Access (CLIA)	• Development of web-based cross-lingual information Access • The query language –Retrieval language pairs are: Hindi -Hindi , English Bengali- Bengali , Hindi, English Marathi- Marathi , Hindi, English Punjabi- Punjabi , Hindi, English Tamil- Tamil, Hindi, English Telugu- Telugu , Hindi, English
Indian Language to Indian Language Machine Translation System (IL-IL MT)	• Development of Machine Translation System from one Indian Language to another Indian language. • The language pairs are: – Tamil-Hindi ; Hindi-Tamil – Telugu-Hindi; Hindi-Telugu – Marathi-Hindi; Hindi-Marathi – Bengali-Hindi; Hindi-Bengali – Tamil-Telugu; Telugu-Tamil – Urdu-Hindi; Hindi-Urdu – Kannada-Hindi; Hindi-Kannada – Punjabi-Hindi, Hindi-Punjabi – Malayalam-Tamil ; Tamil-Malayalam
English to Indian Language Machine Translation system (E-IL MT)	• Development of Machine Translation System from English to Indian languages. • The language pairs are: – English-Hindi – English-Bengali – English-Marathi – English-Oriya – English-Urdu – English-Tamil.
English to Indian Language Machine Translation system with Angla-Bharti Technology	• Development of Machine Translation System from English to Indian languages. • The language pairs are: – English-Urdu – English-Punjabi – English-Bengali – English-Malayalam •Angla-Bharti-II Technology will be used in this project for the development of MT system for the specific language pairs

In the above six consortium projects 28 institutions are participating.

The details are given below:

Institution	Text OCR	Online Handwritin Recognitio n	CLIA	IL-IL MT	E-IL MT	E-IL MT Angla-Bharti)
IIT Delhi	◈ (CL)					
ISI Kolkata	•	•	•			
IIIT Hyderabad	•	•	•	◈ (CL)	•	
IISc Bangalore	•	◈ (CL)		•	•	
Univ. of Hyderabad	•			•		
Utkal Univ. Bhubaneshwar	•		•		•	
MS University Baroda	•					
IIIT Allahabad	•			•	•	
C-DAC Noida	•		•	•		•
Punjabi University, Patiala	•					
C-DAC Pune	•	•	•	•	◈ (CL)	
IIT Madras		•				
CK Technologies, Chennai		•				
Learnfun Systems, Chennai		•				
C-DAC, Mumbai					•	
IIT Bombay			◈ (CL)	•	•	
Jadavpur University , Kolkata			•	•	•	
Amrita University, Coiambatore					•	
Banasthali Vidhyapeeth , Rajasthan					•	
IIT Kharagpur			•	•		
AU-KBC, Chennai.			•	•		
Tamil University, Thanjavur				•		
AU-CEG , Chennai			•			
C-DAC Thiruvananthapuram						•
C-DAC Kolkata						•
IIT Kanpur						◈ (CL)
STQC, C-DAC (for Testing)						
Total	**11**	**7**	**10**	**12**	**10**	

✓ Progress of the consortia mode projects:

1. English to Indian Language Machine Translation System (E-IL) MT

The objective of this consortium project is to develop and deploy a Machine Translation (MT) System from English to Indian Languages in Tourism and Health Domains for the following language Pairs :

English–Hindi

English–Bengali

English–Marathi

English–Oriya

English–Tamil

English–Urdu

For developing the system, various linguistic tools and resources namely Format Extractor, Morph Analyzer, Named Entity Identifier, Parts-of-speech (POS) Tagger, Word Sense Disambiguation, Example Based MT, Statistical MT, TAG Based MT (Parser and Generator), Anal-Gen Based MT (Parser and Generator), Semantic Feature on TAG Trees, Post Processing Tools and Linguistic Resource Management Tools would be developed. The development of the entire system involves various horizontal tasks which are language independent modules and language verticals.

Vertical Tasks

- **Parallel Corpus Collection (V1)**
  Collection and compilation of corpus for the language pairs are in progress. 152 files were distributed among 8 institutions for parallel corpus creation. So far a corpus of 15200 sentences having 2,50,000 words have been created.
- **Corpus Tagging**
  The automatic POS (Parts Of Speech) tagging of the Corpus for annotation purpose is completed and the above-mentioned corpora presently in progress.
- **Grammar Creation**
  The Tree Adjoining Grammer (TAG), AnalGen and Statistical Machine Translation (SMT) Grammatical Formalism have been adopted in EILMT. Grammar creation with TAG for 100 sentences for the language pairs is in progress.
- **Bilingual Lexicon Building:**
  A lexicon of 907 word and phrases for 100 sentences have been distributed among 8 institutes for Bilingual-Lexicon creation. Out of 6 language pairs lexicon creation for English-Hindi is completed. It is in progress for other languages.
- **Parser and Generator Modules**
  Parser and Generator on for English-Hindi language pair are being developed.

Horizontal Tasks

Horizontal task	Institute Name	Status
Morph Analyzer	IIIT –H	This module sent by IIITH has been modified by C-DAC Pune for database connectivity and proper schema. Still two intermediate text files are being used. IIIT-H has to send modified code to read these files from Database
POS Tagger	IIIT-H	POS Tagger has been reimplemented in JAVA by C-DAC Pune and has been integrated.
NER	IIT-B	Modifications required to be done for - module to work with UNTRAINED DATA. - module to work with intermediate data from database instead of Text files
WSD	IIT-B	Modifications required to be done for - the module to work with intermediate data from database instead of Text files
SMT	C-DAC-M	Modifications required to be done for - I/O and intermediate data to be read from database instead of Text files.
EBMT	IISc-B	Not Available
AnalGen	IIT-H	Integration is yet to be done
Collation and Ranking Module	C-DAC M	Yet to start
Evaluation Module		Modifications required to be done for module to work with - proper GUI for Human Evaluation, BLEU and Modified BLEU - complete paragraph
LRMT	IIIT-A	The version sent by IIIT-A does not have features of LRMT like aligned corpus creation, lexicon building and POS tagging.
	C-DAC P	Advanced version has been prepared by CDAC-P including all features of corpus management and lexicon building. This version of LRMT has been distributed among consortia institutes for lexicon creation of their respective languages.
User Log module Input Format Extractor Post processing	C-DAC P	Completed Working fine for .RTF and .HTML.For .XLS and .DOC is in progress Completed

2. Indian Language to Indian Language Machine Translation System (IL-IL MT)

The objective of this consortium is Development of Indian Language to Indian Language Machine Translation system, presented the overall progress of the consortium. He gave a brief background for initiation of consortium mode projects. The systems would be real-life systems, with a given level of translation accuracy, and would be capable of further improvement using machine learning techniques. The language pairs involved are:

- Tamil-Hindi ; Hindi-Tamil
- Telugu-Hindi; Hindi-Telugu
- Marathi-Hindi; Hindi-Marathi
- Bengali-Hindi; Hindi-Bengali
- Tamil-Telugu; Telugu-Tamil
- Urdu-Hindi; Hindi-Urdu
- Kannada-Hindi; Hindi-Kannada
- Punjabi-Hindi, Hindi-Punjabi
- Malayalam-Tamil ; Tamil-Malayalam

The progress of the consortium as presented by the consortium leader is as follows:

S. No.	Horizontal Task	Status
1.	Tokenizer	Done
2.	POS Tagging & Chunking Engine	Done
3.	LWG	Done (to be validated)
4.	Morph Analyzer Engine	Done
5.	Generator Engine	Done
6.	Lexical Disambiguation Engine	Delay
7.	Named Entity Engine	(CLIA consortium to give)
8.	Dictionary Standards	Done
9.	Corpora Collection Standards	Done
10.	Corpora Annotation Standards	Done
11.	Evaluation of Output Comprehensibility	Delay
12.	Testing & Integration	Done
13.	Transliteration	Delay
14.	Transfer Engine	Delay

Sl. No.	Task	Bengali	Hindi	Kannada	Malayalam	Marathi	Punjabi	Tamil	Telugu	Urdu
1.	POS Tagger	Done	Done	Done	Done	Done	Done	Done	Done	Done
2.	Chunker	Done	Done	Done	Done	Done	Done	Done	Done	Done
3.	Morph Analyzer	Done	Done	Delay	Done	Done	Done	Done	Done	Delay
4.	Generator	Delay	Done	Delay	Delay	Delay	Delay	Delay	Done	Delay
5.	Named Entity Recognizer	---	---	----	------	------	----	-----	-------	-----
6.	Bilingual Bi-directional Dictionary	Done	Done	Delay	Done	Done	Done	Done	Done	Done
7.	Transfer Grammar Component	Delay	Delay	Delay	Delay	Delay	Done	Delay	Delay	Delay
8.	Corpora Collection	Done	Done	Done	Done	Done	Done	Done	Done	Done
9.	Corpora Annotation	Done	Done	Done	Done	Done	Done	Done	Done	Done
10.	Evaluation									

3. **English to Indian Language Machine Translation (E-IL) System project with Angala-Bharti Technology:**

The objective of this consortium is to develop English-Indian Language Machine Translation with Angla-Bharati presented the overall progress of the consortium. The language pairs involved are: English-Bangala, English-Punjabi, English-Urdu and English-Malayalam. The progress of the consortium as presented by the consortium leader is as follows:

Task -1	Familiarization with AnglaBharati System
Task -2	Familiarization with AnglaBharati Lexical data-base structure
Task -3	Familiarization with tables/symbols used in AnglaBharati requiring modifications
Task -4	IL entry & display units integration. This task has been completed for English to Bangla/Malayalam/Punjabi/Urdu. IITK romanized code has been worked out for each of the language. Entry is possible in both the sch emes (IITK and inscript keyboarding). Display using Unicode also has been implemented.
Task -5	IL paradigm file generation
Task -6	IL Morphological Synthesizer The tasks are completed for Bangla, Malayalam and partially completed for Punjabi and Urdu . This may require revisit when the text generator part is completed.
Task - 7	English- IL Transliterator This task is completed for Bangla, Malayalam , Punjabi and Urdu .However it is not thoroughly tested for Malayalam , Punjabi and Urdu .
Task - 8	IL symbol mappings as per AnglaBharati engine This task is completed for Bangla, Punjabi and Urdu. However further testing is needed for Punjabi and Urdu languages . For Malayalam 90% task has been completed .
Task - 9	IL Text Generator (partial)

The progress of various language verticals are as follows:

Bangla : Text generator for Bangla simple sentences (affirmatives , imperatives, command, request, let) has been implemented based on PLIL structure for the above mentioned types of sentences.

Malayalam : This task is completed for simple sentences . However, further testing is needed to find gaps.

Punjabi : This task is completed for simple sentences . However it requires further testing.

Urdu : This task is completed for simple and compound sentences . However further refinements are in progress.

The status of corpora collection is as follows:

Health:

- A corpus has been collected from the book (title : "Where There is no Doctor") .
- 14.3 mb of Corpora from hse.gov.uk and drugs.com
- About 1 GB of Corpora from drugs.com
- 71 mb of Corpus Collected from who.int,en. wikipedia.org
- 10.9 mb of corpus collected from hpathy.com and other sites
- 102 kb corpus collected using pathologytraining.org & other sites
- 662 kb of corpus collected using radiologyinfo.org and other sites
- 5.28 mb of corpus collected using google.co.in and other sites

Ayurveda

- Vagbhata's Astanga Hrdayam – 3 Volumes (1733 pages)
- Carakasamhita – 4 Volumes (2240 pages)
- Susruthasamhita – 3 Volumes (1389 pages)
- Ayurveda - 500000 words

Homeopathy

- Materia Medica- Robin Murphy (1893 pages)
- Home guide to Medical Emergencies – Dr. Henry & Lawrence Galton (204 pages)
- Materia Medica and Repertory–William Boericke (1986)
- Homeopathy - 52 lakh words

Tourism

Kerala Tourism - 38 lakh words from14 Handbooks/ Leaflets.

4. **Development of Cross-lingual Information Access**

The objectives of the project are:

To Develop a portal where,

(1) A user will be able to give a query in one Indian Language and

(2) The user will be able to access documents available in

(a) The language of the query and
(b) Hindi (If the query language is not Hindi) and (c) English.

The query language –Retrieval language pairs involved are:

- Hindi - Hindi, English
- Bengali - Bengali, Hindi, English
- Marathi - Marathi, Hindi, English
- Punjabi - Punjabi, Hindi, English
- Tamil - Tamil, Hindi, English
- Telugu - Telugu, Hindi, English

The progress of various Horizontal and vertical tasks are as follows:

* Horizontal Tasks

Input Processing

Stemmer

- All the language verticals have developed their respective language Stemmers.
- These stemmers have been integrated with the Nutch Morph Anlayzers by writing suitable plug-ins.
- Monolingual retrievals have been independently tested for each language
- Hindi language identification is being addressed.

Multi Word Expressions

- The guidelines for identification of MW are under discussion at IITB and will soon be finalised. Until then it has been suggested to derive the MWL by identifying the top 2000 or so most frequent bi-grams / tri-grams.
- IITB has around 2000 MWE, Bangla has around 600 MWE, Tamil has around 500-600 MWEs.

Dictionary

- While all institutes are working on H-IL mapping primarily, IITB is working on Eng-Hin linkage. This would enable the E-IL linkage to be made automatically for the common synsets.

WSD

- The WSD developed by IITB requires sense-marked data for testing. A tool (software interface) has been developed by IITB that helps to build sense-marked corpora.

Search

Indexing

- IIT Kgp has indexed 10,000 pages of English
- IIIT H has indexed 21,000 pages of Hindi. This includes around 200 documents in each of Tourism & Health domains.
- Bangla - 25 URLs - 3000 documents of News, Health and Tourism
- Tamil - 25 URLs
- Telugu - 40,000 pages
- Punjabi - 25 URLs - 17000 pages
- Marathi –200000 pages

cML-Text Converter

- The first version of the engine is ready. The software extracts the fields and body, but does not identify paragraphs and blocks in this version. It is ready to be integrated with Nutch.

Document Processing

Information Extraction

- The IE Template is ready and the basic IE engine can be demonstated.

Output Generation

Snippet and Summary generation are ready to be demonstrated.

Evaluation

Corpora

ISI Kol has initiated talks with TOI and Hindustan Times for permission to use their multilingual corpora, which could be used by all the institutes.

Status of Corpora

Language	News	Source		Tourism	Health
Bangla	> 2,00,000 docs	Anandabazar Patrika	2004-2007		
Hindi	1,00,000 docs				
Marathi	~ 2,00,000 docs	Maharashtra Times, Sakal.com Webdunia Marathi Manogat	2002-2006	2,40,000	2,60,000
Tamil	10,000 docs	Dinamani, Dinamalar, Dinathanthi, Dinakaran, BBC Tamil	2007		
Telugu	25,000 docs				

Topics

Topic Translation has been completed by all the six language verticals.

The final set of 30+50(+15) topics for evaluation are ready. 30 topics for training and 50 topics for testing. These have to be translated into six languages.

UNL

Monolingual information retrieval is working for Tamil.

Testing and Integration

Test report and Integration Plan document has been prepared

5. Printed Text OCR system:

The objective of this project is to develop and end-to-end OCR system for printed Indian scripts, for possible conversion of legacy printed documents into electronically accessible format. The scripts are Bengali, Devanagari, Malayalam, Gujarati, Telugu, Tamil, Oriya, Tibetan, Nepali, and Gurumukhi. The progress of various Horizontal and vertical tasks as are as follows:

Documents: Software Requirement, Design and testing Document have been prepared.

Scanned Document Database: A collection of scanned document images, for Gujarati,

Hindi, Tamil and Kannada were released by IIIT Hyderabad in July, 2007. Another sample Text Image corpus consisting of scanned text images for Telugu, Gujarati, Kannada, Bangla, Tamil, Malayalam and Hindi script was released by IIIT Hyderabad in October, 2007. The documents are scanned at 200dpi, 300dpi and 600dpi.

Codes released : Alpha version of the pre-processing tools and segmentation library have been released.

A list of codes released is given below.

1. Pre-processing Tools
2. Document Image Segmentation Tool
3. Content Classification and Labeling Tool
4. XML I/O Library

System Integration:

This work is in progress by CDAC, Noida. Currently they have integrated the Pre processing Tools and Segmentation Tool provided by IIIT, Hyderabad, IIT Delhi and ISI, Kolkatta. CDAC, Noida has made use of Broker Architecture for the integration of the API's to enable autonomous and heterogeneous system to share information while maintaining autonomous control.

User Interface:

The presentation engine will be used to display the final output of the OCR system which is obtained in the form of an XML encoded document. The XML output document has the pre-processing details, layout structure of the document, the various segmented regions and their categorization (text, graphics, pictures, etc.). The presentation engine will make use of the XML-encoded output and display the electronic document with the document-layout information preserved to the best possible extent. This work is also in progress by CDAC, Noida and a version has been released for use.

6. On-line Handwriting Recognition System

The progress of the consortium is as follows:

- Collection of Data using Comprehensive Word / Character List
- **Tamil Script:** A set of 93 words covers all the possible characters and symbols in Tamil. Data has been collected from 100 native writers.
- **Kannada Script:** A list of 140 words covers every possible basic and derived character. Again these words have been collected from 80 different native writers of Kannada.
- **Bangla:** Samples of handwritten words of Bangla are being collected on the basis of a list of 687 Bangla words. This database is expected to have representation of more than 98% of character occurrences of any standard corpus.
- **Malayalam and Telugu :** A list of 120 words for Malayalam is collected each from 10 writers and 150 words each from 100 Telugu writers have been collected.
- **Devanagari :** A list of 800 words were generated that covers all the possible symbols in the language. Data has been collected from 30 writers.

* Input Spatial and Time Resolution: Both for training and testing, handwritten data must be captured at or above 500 dots per inch and 100 samples per second. Data has been collected using the GENIUS pad, with a sampling rate of 160 points per second and resolution of 1000 DPI and / or TABLET PC for

Tamil.

- Annotation

Requirement Specifications for Annotation tool has been created, to segment the collected data at different levels. Levels can be page, paragraph, lines, words, or characters. The position of the Unicode values in the annotated page should be as close in correspondence to the location of the actual strokes in the digital handwritten page.

* Annotation Tools

IIIT Hyderabad and ISI Kolkata have created Annotation toolkits currently used for Devanagari, English & Bangla for the purpose of annotation of words, characters and strokes.

* Features Used for representing the characters

Spline Fitting, chain Code Representation, Curvature-Tangent-Height Representation, and offline Features such as, Projection Histograms in X and Y directions, Mean and variance of pixel values in different sub-windows, Number of horizontal and vertical crossings at different positions and Distance-transform representation and its Eigen representations.

* Architecture for the Recognition Engine

Architecture has been created for the recognition engine, discriminating between language dependent and language independent modules.

* Classification of Handwritten Characters

Both single classifier and multiple classifiers are being tested. Further, a hierarchical classification strategy is being adopted. The classifier design is in its initial stages, and experiments have been started on various strategies to evaluate their performance on the data set that is being generated. The preliminary recognition performance obtained for various languages, using limited data sets are:

Language	Mode	Stroke/Character	Feature	Classifier	Accuracy
Tamil	Writer dependent	characters	Equiarc length coordinates	Subspace	93%
Tamil	Writer dependent	characters	processed coordinates	DTW	96%
Telugu	Writer Independent	Strokes	Position, curve length, octant based	SVM	94.5%
Hindi	Writer Independent	Strokes	Spatiotemporal and spectral	SVM	94.5%

Current Status :

All the above Consortia mode projects are under implementation and alpha versions of the software systems are going through software engineering and linguistic testing.

Contributors :
Ms. Swaran Lata, Dir., TDIL, DIT
Dr. Somnath Chandra, Jt. Dir., TDIL, DIT
Mr. Vijay Kumar, Jt. Dir., TDIL, DIT

Annexure -1

ROADMAP FOR TECHNOLOGY DEVELOPMENT FOR INDIAN LANGUAGES (TDIL)

FROM TECHNOLOGY DEVELOPMENT TO APPLICATIONS

Department of Information Technology
Ministry of Communications & Information Technology
Govt. of India
Electronics Niketan
6, CGO Complex, Lodhi Road
New Delhi – 110 003

Contents

Annexures

1 INTRODUCTION

1.1 Background

Language technology has reached a stage today where it has developed a potential to generate utility applications, benefiting the masses, which will enable people to access and use IT solutions in their common language. These utilities can help processors like translation to and from Indian languages, scan and OCR Indian language contents in physical form, handle databases, access Internet, handle e-mail in their own languages. Besides, these activities can help illiterates by the use of text-to-speech and speech-to-text utilities, giving proliferation to IT in rural segments as well. High-End tools like sophisticated search engines and content creations can help take Indian languages at international level. Indeed, the Digital Library of India can be made more accessible and can be used to provide Indian literature to the world and rural masses, with speed and affordable costs.

Localisation of applications focusing on display and keyboard handling in their own language have also been a key component, although it mainly pertains to customization, development and deployment. It also, in a way, relates itself with standardization.

Keeping in view of all the above, the Department of Information Technology (DIT) has thought of encouraging users and developers of Language Technology solutions by providing free of cost, certain basic information processing tools like fonts, open office, e-mail client, internet browser, dictionary, conversion utilities, etc., which will motivate users to use them to solve their basic problems and help developers to build advanced solutions. This will definitely boost up and leapfrog Indian language technology development and their deployment in a very fast way.

This report formulates short-term and long-term mission plans for these activities with appropriate budget estimates and duration.

1.2 Mission Statement

Enabling masses to build knowledge society.

1.3 Goals

For large-scale benefits of this technology to be realized, mass applications need to be developed as quickly as possible. This development should go along with a long-term vision. This area also needs to attract the attention of a lot more researchers and students in India, than it is currently doing.

One way to energize this area is to set visible missions. These should have definite deliverables in the short term, in the next one to two years. Along with this, the long-term vision should be set, so that the short-term and the long -term dovetail with each other.

One can also set "grand challenges", which would be futuristic. These would serve to excite the imagination of users and researchers alike. The grand missions would work to develop advanced technology. This would have the potential to establish India as the most important language centre internationally. The world leadership achieved by the Indian IT industry for service-oriented work can also be achieved in language technology area, wherein work for other languages of the world would come to India and use the technology developed here.

Following goals need to be achieved under the mission plan:-

i. Set up mission mode projects for mass uses
ii. Dovetail short-term applications with a long-term vision
iii. Aid to develop technology of international repute so as to help India emerge as a world leader in language technology
iv. Provide necessary infrastructure and mass proliferation of the fundamental technologies to take a feedback of prototype and similar applications
v. Devise a plan for developing manpower to work in Indian language industry government, public and private, so that as technology use increases, manpower would be available not only to further technology but also to provide necessary direction to the mission implementation.

1.4 Applications

The technical categorization of applications and utility needs by the masses has been taken into consideration to workout following applications, to meet the challenge under this mission. These applications are:-

i. **Mass access to information.**
   * Telephone based (voice-enabled) information access.
   * Information Kiosks with multi-modal interface.
   * TTS on SMS on Mobile Platform
   * Domain Specific Dialogue System

ii. **Cross lingual access**
   * Web Based Cross Lingual Information Retrieval
     - Machine Translation among Indian Languages
     - Machine Translation : English to Indian Languages
     - Machine Translation : Indian Language to English

iii. **e-Content Creation**
   * Scanning books in Indian languages to generate electronic text:OCR
   * Communicating over PDAs in handwriting

iv. **Localisation**
   * Localisation of Middleware
   * IDN & e-mail id in local languages
   * Localisation of content (English --> Indian Language)

These applications have been considered of mass utility and have been technically evaluated in respect of technical strength of the committee, institutions working in this area in India and a forward vision for technology development. In view of combining efforts of all such institutions to aligned development in specific directions, this document suggests that requirement for proposals (RFP) should be developed in some details. It has been done and added as annexure in this report.

1.5 Related Technology Development to Generate Applications

In view of the applications mentioned above, the underlined technologies have also been gone through, although some fundamental work has already been done, and various methodologies in technology development have a version of similar applications, it is needed to strengthen these technical developments as well. These technologies are:-

i. Speech processing
   * Speech recognition
   * Speech synthesis

ii. Natural Language Processing (NLP)
   * Machine translation (MT)
   * Information Extraction & Retrieval (IR)
   * Semantic Search

iii. Optical character recognition (OCR)
   * Indian Language OCR
   * Indian Language on-line Handwriting Recognition (OHR)

iv. Localisation
   * Transliteration amongst Indian languages
   * Standardization in localization benefiting e-governance
   * Localisation of Middleware
   * IDN & e-mail id in local languages

While the mission mode projects are formulated with definite deliverables, it becomes essential to observe the development in the underlined technologies, which can be made available for use across host of applications. Secondly, it is the experience worldwide that products out of these technology developments need continuous research and improvement so that their accuracy and functionalities can continuously improve. For this purpose, testing & evaluation also become essential. Therefore, the technology should be so developed that it remains within the ambit of research for its constant enhancement.

2. SHORT-TERM GOALS

As per the Mission Document, it is expected that the following fundamental tools be made available free of cost to the end users and to developers for the high-end products, ultimately benefiting masses. This tool kit will be called as Basic Information Processing Kit (BIPK). It will consist of :-

a) Fonts for local language
   * TTF
   * OTF

b) Open Office for local language

c) Word Processor in local language

d) Bi / Tri-lingual Dictionary

e) e-Mail Client

f) Internet Browser

g) Code / file conversion utiliteis

In addition to the above, following additional tools can also be given out wherever available from resource centres, coilnet centres, educational institutes and R & D institutes like C-DAC.:-

* Simple Machine Translation System and tools such as Morph Analyser, Spell Checker
* Transliteration Tool
* OCR in local language
* Text to speech System
* Speech to text System
* Text Corpora
* Speech Corpora
* Content Creation Tool

Annexure V gives the availability of these tools in prototype. These are available from various resource centers and other agencies where DIT has funded the activities under TDIL programme.

3. STRATEGY FOR TECHNOLOGY DEVELOPMENT

The mission mode projects with specific deliverables can be successful if definite strategies are applied. For research, funding of the project can be made through normal mechanism, but for mission mode projects, funding has to be made available with specific action plans in mind and with clear-cut input, output and outcome. It is suggested that the following parameters may be considered while handling the strategy of funding:-

i. Separating Language Specificity from Technology

So far, language has been considered as vertical and various products and technology development for that language has been sent as projects for funding. After having gone through such a drill, and creating awareness, it is now evident that a level of maturity has been achieved to consider technology development as a vertical, under which different languages can be embedded. The time has come that various engines can be designed and developed which can be tuned to specific languages later. This will not only maintain harmony amongst all Indian languages but also will provide a standardized way and path for technology development. For such projects, one needs to design generic engines and develop appropriate data for different languages. While the task of data preparation can be distributed amongst language specific groups, the development of language technology engines can be done by specific groups. The groups developing engines, need to work closely with language data formulators so as to integrate language specific aspects into the technology being developed. It is also suggested that such engines be tried with 3 or 4 languages initially and suitable mechanism should be devised to tune them with other Indian languages.

(Note: The term 'engine' is used here to refer to software implementation of the algorithms, which can be trained or adapted using data from different languages. Also note that although the engine should be separable from the data, its development must be tied with at least one language to monitor its development. Moreover it may be made a part of the mission mode project so that the end results become visible at an early date.)

ii. Setup Language Technology Groups as Consortia

The language technology development fruits up if linguists, language experts, language engineers and technical engineers shake hands properly. The prevailing atmosphere in India is very conducive for such groups working together, however, it needs conscious nurturing at the level of funding agency. The group working on languages, shall prepare data and train the engines using the data. It would mean that for such groups, scientists as well as linguists are needed. For example, for preparation of the annotated corpora and lexical databases, care need to be taken to link them with technology development. Consortia formulation would take care of this link. Looking at Indo-Aryan and Dravidian structure of Indian languages and in view of multiplicity of languages and wide geographical spread, it is proposed to set up a consortia per cluster of languages.

Technology Engines will be trained with data for specific languages.

iii. Allow Alternative Approaches in Technology Engines

Technology development, fundamentally, is a research activity and therefore, it needs to be

nurtured by permitting alternative approaches in formulating technical engines. However, with the work done so far and the paucity and limitations of the funding pattern, it is suggested that one or two alternative approaches be funded under the mission. Other approaches may be kept open for research at institutions such as IITs.

Some important but difficult components of the technology engine may be opened up for competition and contest. For example, in a machine translation engine, word-sense disambiguation or verb frame selection components are hard to build with good accuracy. The competition may bring out new approaches applied to a difficult problem (by older or newer groups). The competition will also lead to a more rapid development.

iv. Evaluation of Technology Engines and Their Components

For the success of a mission, it is important to identify set up measures for the evaluation of technology needs and / or their importance of sub-components. This allows technology development to be monitored leading to formulation of end user products.

The measures, in fact, become the basis for evaluating alternative approaches of alternative groups working on the same problems. The best approaches can be selected and further refined. This has been used successfully in advanced nations to make rapid advances in this field.

The monitoring committee must identify important components, recommend the promising competing proposals, and then evaluate the results. In other words, much greater amount of technology inputs are needed to develop and test components, and integrate into end-user systems.

v. Evaluation of End-User Systems

Evaluation criteria need to be developed for the end-user systems, using evaluation of each component sub-system. Earlier evaluation mechanism was based on specifications for user acceptability in a deployed application The proposed evaluation mechanism is based on evaluating component sub-system.

vi. Mechanism to Seed New Groups

Much larger number of groups are needed to handle different facets of language technology than currently available in the country. New groups can be given smaller tasks (and smaller funds) to begin with. If they deliver as judged based on quantitative evaluations, they can be given larger tasks and funds.

This method would set up a continuous process of bringing-in and weeding out approaches and groups.

vii. Encouraging Basic Research

One way to seed new groups is to set aside some funds for basic research. A field advances when researchers with new ideas join and contribute. Based on the ideas, new approaches emerge. A major problem being faced by NLP research in the country today is the acute paucity of research funds. Many bright individuals who start their career in NLP or related areas move to other areas since they find no funds are available in NLP, speech, etc.

As part of a promotional policy to draw more researchers into language technology, certain part of the funds should be set aside for promoting basic research, irrespective of whether it would directly contribute to a known technology. The level of funding to individual basic research projects not linked to coordinated technology development effort should, however, be kept low, say, a ceiling of Rs.10 lakhs for 2 years.

viii. Encouraging Human Resource Development

The area of language technology has been suffering from lack of trained man power, because, first, the students of computer science have been seeing it as a data preparation activity and not as a technology development activity, and second, jobs had not come up in this area in Industry.

Fortunately both the above are changing today. Jobs are coming up in industries such as HP Labs., Google, Tatas, Reliance, IBM, Microsoft etc. The numbers of jobs are still small but they are high paying ones. They are growing steadily in number.

Language Technology students face many challenges in areas such as machine learning, speech and OCT. This trend needs to be augmented by encouraging innovation of intensive projects and the formal education programmes such as M.Tech. Ph.D. etc. in the field of language technology.

4. OTHER IMPORTANT POINTS

For making a mission success, it is felt that technology development be open amongst the concerned developers and consortium members. Therefore, it is suggested that the following activities be started:-

i. Handling Corpora and Sharing of Programs

A long-standing problem on the Indian scene has been the lack of public availability of resources (whether data or programs). Even when such resources get developed from project funds, they do not become available to other funded projects even when the project is from the same funding agency. The work is done again, usually losing valuable time and money.

Whereas a separation between technology engines and linguistic data can help improve the situation, proactive steps need to be undertaken.

A consortium needs to be set up which can distribute data and programs for research use. For linguistically annotated data and lexical resources, certain amount of technical management is needed. Therefore, it is best to locate it within an academic institution. World's most successful consortium runs at an academic institution (Linguistic Data Consortium at University of Pennsylvania), and undertakes not only the distribution of data but also evaluation, clean up, and creation of data resources.

Projects funded by DIT as well as other ministries, may contribute their data to the consortium. Industry may also take the data for research use from the consortium on payment basis. For commercial use, it will have to negotiate directly with the "owner" of the data.

ii. Handhold Support to End Users

While prototypes and product versions zeros are given to the users, users can intelligently use them and generate a feedback. Since such feedback has been generated by the end users, it requires to be categorized and correlated with technology development. This report suggests that a group be set up exclusively for this activity.

PART II: ADMINISTRATIVE AND FINANCIAL

5. ADMINISTRATIVE ISSUES

5.1 Mission Mode Projects

i) Short-Term Goals – Preparation of Basic Information Processing Kit in Indian Languages

- Fonts (TTF and OTF)
- Open Office in Local Language
- Word Processor
- Bi/ Tri-Lingual Dictionaries
- e-mail client
- Internet Browsers in Local Language
- Code Conversion Utility
- Beta or zero versions of
  - » Machine Translation Tools
  - » Transliteration Tools
  - » OCR for printed material
  - » Text to Speech converter modules
  - » Other related products available if any

ii) Medium/ Long-Term Goals

- Mass Access to Information
- Cross-Lingual Information Retrieval
- E-Content Creation
- Localization
  - » Transliteration amongst Indian Languages
  - » Standardization in localization benefiting e-Governance
  - » Localization of Middleware
- Human Resource Development
- OCR & OHR for printed material
- Speech Processing
- Speech Synthesis
- Speech Recognition
- Machine Translation Tools
- Mechanism to develop e-mail ID in Indian languages
- Text Processing analysis Tools

Following major applications are launched in mission mode project:-

i) Preparation of BIPK short-term goal
ii) Upgradation for products launched in short-term goal
iii) Telephone / Voice Based Information Access
iv) Machine Translation : (a) English to Indian Language & (b) Among Indian Languages
v) Cross Lingual Information Retrieval with Eng IL
vi) Optical Character Recognition (OCR) & on-line Handwriting Recognition (OHR) in Indian Languages
vii) Transliteration amongst Indian Languages

This report includes Request For Proposals (RFP) for iv, v and vi above. For telephone based information access project at iii, joint funding, as requested by Department of Science and Technology, Ministry of Science and Technology could be considered. However, work related with I and ii above i.e. preparation and upgradation of BIPK and transliteration amongst Indian Langauges, can be given to different units of C-DAC, which have been the nodal agency for integrating the products, as per short-term goal and releasing CDs in different languages so far. These RFPs are indicative and minor changes may be incorporated once expert groups are formed accordingly and discussions are held with developers and other agencies.

5.2 Project Implementation Strategy

For each of the mission mode projects, it is expected that a number of institutions would get involved. This is inevitable for two reasons. First, the technology involved is very complex and no single group would be in a position to deliver. Second, many languages would be involved. This would make it very difficult if technology for all of them had to be developed at a single place.

What this means is that the technologies and modules produced by different groups need to be integrated and tuned. This requires some common standards for representation of information. These would have to be arrived at through common scientific meetings.

It is proposed that while design of engine, independent of language, happens separately and its instantiation for 4-5 languages takes place, the development of language data required, can start simultaneously, for all 22 Indian languages. For each Mission Mode Project, special coordination team has to be set-up. The team will consist of 3 key persons:

- Chief Coordinator,
- Chief Architect, and
- Chief Technologist

The Chief Coordinator will be in-charge of the overall monitoring and delivery of the project. Financial requests of individual institutions have to go through the Chief Coordinator to MCIT for coordination to be effective.

Chief Architect will be responsible for taking scientific decisions pertaining to common representations or the overall architecture of the system. He would help setup the standards and will also handle change-requests, version control during the implementation.

Chief Technologist will be in-charge of system integration, testing, software engineering, documentation, and maintainability of the software produced by the project. It will be desirable that a person from industry gets involved in this.

5.2.1 Putting the Institutions Together

Key task for MCIT would be preparing detailed RFPs as per recommendations, issuing them and processing the proposals received.

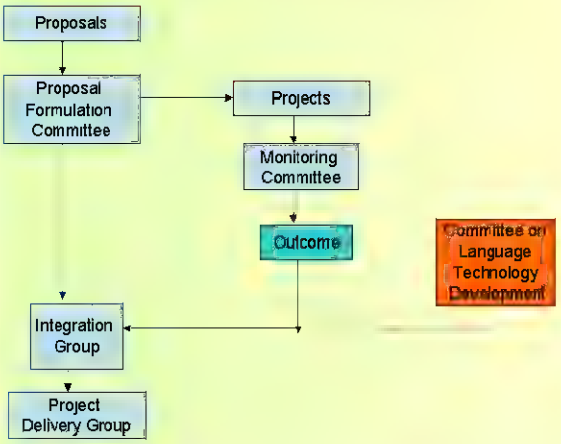
A meeting of the prospective institutions who submit a proposal or are likely to submit a proposal must be called to put effective coordination mechanism in place, to do a division of major tasks among institutions, to propose languages to be covered in the first phase, and to set up expectations and delivery.

The system integration is an extremely important part of an application. The components or modules developed by a group have to work together with modules of other groups. A system integration group distinct from the module development group would have to take up the task of integration. It would be responsible for running and testing the application system.

While the above are integral part of the implementation groups in case of such a mission-mode project, the funding agency (MCIT) may set up a Project Review & Steering Group, in which International expert(s) may also be included. This would add professionalism and international perspective to the project.

In case of Mission Mode Project, Project Review and Steering Group (PRSG) needs to be set up. This group will have 1 member from the PRSG along with other national and international experts.

Project Implementation Mechanism Proposed



Project Monitoring Board will have a member from Committee

Time scale for processing of RFP:

- Day O: Issue of an RFP (Expression of Interest with brief proposals to be submitted within 15 days, and a detailed proposal within 30 days)
- Day 15: Expression of interest with brief proposals received from agencies
- Day 21: Meeting of agencies who have expressed interest (to be initiated by MCIT)
- Day 30: Updated detailed proposal(s) received by MCIT (including possibly a joint proposal)
- Day 45: Processing of proposals to be completed by MCIT
- Day 60: Projects granted against proposals

5.3 Project Implementation Guidelines

- All projects must have a two-page summary that will be publicly available. The summary will list the modules that the project is developing internally as well as modules that are being re-used/ shared by other projects. These must be given along with release times for the indicated modules for alpha, beta version and so forth. The projected accuracy of the modules must be listed along with evaluation criteria.
- All TDIL-funded projects will make internally developed modules available to other TDIL-funded projects at the specified release times. For early releases, the binary will be distributed.
- Source code will be given to MCIT to allow further development in subsequent phases. PI's from other mission projects who have planned their projects around a specified module may attend evaluation meetings to track progress and accuracy.
- Developers of core engines will make their engine available at early stages also, and on a timely basis, as specified in the project write-up.
- Release times: Alpha-versions of all projects will be released within six months of commencement, and code will be accessible to developers of other mission-mode projects.
- Bug tracking: the System Coordinator will enable bug-tracking through a system such as Bugzilla. Follow-up on bugs will be scheduled on maintenance or release basis, and notified through the system. All TDIL projects with dependency on internal modules may post bug reports.

Release Times :

Alpha-versions : Within six months of commencement, and code will be accessible to developers of other mission-mode projects.

Bug Tracking : The system Coordinator will enable bug-tracking through a system such as Bugzilla. Follow-up on bugs will be scheduled on a maintenance or release basis, and notified through the system. All TDIL projects with dependency on internal modules may post bug reports.

5.4 Short-Term Goals

The Basic Information Processing (BIP) Kit consisting of Fonts, Open Office for local language, Word Professor, Dictionary, Conversion Utility, relevant standards, e-mail

client and Web Browser would be launched in all Indian languages. For some Indian languages, OCR, Text-to-Speech and Simple Machine Translation System may also be included. The basic work has been done majorly through agencies like C-DAC, resource centres and other educational institutes with funding through TDIL programme of DIT. These products will be tuned to convert them to usable products and would be provided free of cost to the end user.

Identification, test & evaluation and consolidation of available fonts, tools / technologies from industry, academia and research institutions : Development of necessary user interfaces; and packaging into user-friendly tools and products for earliest launch of Basic Information Processing Kit for Indian languages will be carried out by C-DAC. These will be available on CDs for free to use and also web-downlodable. After launch on-line, user support will be available through the data center and TDIL website maintained by C-DAC.

Schedule of launch of IL fonts and software tools:-

*	Language :	Date :
	Tamil	15 Apr, 05
	Hindi	20 Jun, 05
	Telugu	12 Sep, 05
	Punjabi	24 Sep, 05
**	**Language :**	**Tentative Date:**
	Urdu	Oct., 05
	Marathi	Oct., 05
	Bangla	Nov, 05
	Kannada	Dec, 05
***	**Language :**	**Tentative Date**
	Malyalam	Jan, 06
	Oriya	Feb, 06
	Assamese	Mar, 06
	Gujrati	Apr., 06

For other Indian languages, TDIL team of DIT will take a stock of availability and quality of the products and convert them to usable products in one year's time.

Institutions already working in these areas with or without funding from Government and ready to give commitment for developing usable products in time-bound manner, need to be supported.

5.5 Medium to Long Term Research

Besides working on mission mode projects one should identify areas and technologies which will be needed after 2-3 years for either opening up new applications or to enhance the effectiveness of current applications. A smaller amount of funds may be invested in developing such technologies by promoting them three suitable R&D projects. Some example areas:

1. Text Processing
   - Tree banks in Indian languages (ILs)
   - Sentential parsers for ILs
   - Semantic analysis -verbs and its arguments
   - Word sense annotated Corpora for ILs
   - Word sense disambiguation for ILs
   - Discourse annotated Corpora for ILs
2. Speech Processing
   - New approaches in text to speech systems
   - Exploration of new approaches for speech recognition in ILs
   - Annotated Speech Corpora for ILs
3. OCR and OHR
   - New approaches needed for OCR in ILs to reach high accuracy
   - Special approaches might be needed for OHR for ILs
4. Cross-Lingual Information Retrieval
   - Information Modeling & Extraction
   - Document Summarisation
   - Semantic Indexing
   - Simple Machine Translation
   - High Quality MT System
   - ILs on UNL (Universal Networking Language) of United Nations
   - Integrating different MT approaches
5. Machine Translation Tools
   - Translation memories
   - Translator Aids
   - Dictionaries / Lexicons
   - Machine Translation Test Beds
6. Standardization
   - Unicode, TMX, Web internationalization
7. Transliteration amongst Indian Languages

More detailed long-term plan ought to be prepared in due course.

5.6 Promoting New Groups, Maintaining Continuity of the Old

To draw excellent researchers into the area of language technology, one should consider funding proposals from 'extra-ordinary' researchers even when the proposal does not directly link up with TDIL plans. In such cases, however, only a small sum of money be committed, say, projects of Rs.5 lakhs per year - with a total ceiling of Rs.10 lakhs per year. The intention is to build interest of individual researchers with excellent track record, but at the same time encouraging them to work on problems of greater interest to TDIL and thereby get larger grants.

There is an associated issue of nurturing and maintaining continuity of existing groups that have done good work and completed a project successfully. After the project has completed, if there are long gaps of several months to a year before a new project is given, the scarce manpower at the centre is lost! When new project is given, the faculty member starts from scratch. Usually, he leaves this area and moves to another research area with a better funding.

"Bridge support" should be given by MCIT against well defined deliverables, limited to Rs. 10 lakhs for one year only.

5.7 For monitoring the project, experts from institutes such as C-DAC, IIT, IIIT, IISc and industries will be identified and they will be responsible for steering the project ensuring delivery to the masses.

6. FINANCE

6.1 Mission Mode Projects

Preliminary budget estimates of the RFPs covering 3 to 4 languages in each case is as follows:

S. No.	Mission Mode Projects (first phase)	Duration (years)	Cost (Rs. crores)
1.	Telephone Based Information Access	2.5	7.00
2.	MT : (a) E - ILs	2.0	4.00
	(b) IL - IL	2.5	3.00
3.	CLIR	2.5	7.00
4.	OCR & OHR	2.0	1.00
	Total		**22.00**

The necessary technology engines would be fully developed after about 1 year of the above activity. At that time 8-10 additional languages can be added. (Eventually all 22 official languages can be covered.) However, the development of language data required for all official languages, can start simultaneously.

Mission Mode Projects (later phase)	Phase	Duration (Years)	Cost (Rs. crores)
Cost of additional 8-10 languages	(Year 2 and 3)	2 years	10.00
Cost of additional 8-10 languages	(Year 4 and 5)	2 years	10.00
Total			**20.00**

Costs above are based on projections from some of the mission projects for which detailed costing has been done.

6.2 Other Activities

Other projects include (a) Specialised HRD in Language Technology, (b) Capacity Building at Institutions focusing on technologies & resources for Indian languages in the region, (c) Basic R & D projects plus bridge projects.

	Non-Mission Mode Projects	Phase	Duration (Years)	Cost (Rs. crores)
a)	Manpower Development Programmes	Years(1 to 5)	5 Years	12.00
b)	Capacity Building for Language Technology and Resources	Years(1 to 5)	5 Years	14.00
c)	Basic R&D projects plus bridge projects	Years(1 to 3)	3 Years	4.00
Total			**5 Years**	**30.00**

6.3 Roll out of Indian language technologies : Tools and Resources with zero versions and subsequent improved versions

Non-RFP Project	Duration	Estimated Budget (Rs. In crores)
Launch of Fonts & Software Tools	**2.5 years**	**15.00**
Total		**15.00**

SFCs for (a) Roll-out of Language Technologies, (b) Capacity Building, and (c) Manpower Development programmes may be prepared & processed by the Ministry of Communications and Information Technology.

6.4 Total of the above costs are summarized below.

Summary	Duration (Years)	Cost (Rs. crores)
Mission mode projects	5 yrs	42.00
Other projects	5 yrs	30.00
Roll-out of LTs	2.5 yrs	15.00
Net Total		**87.00**

Annexure I

General Guidelines for Proposal

1 Proposal Details

All projects must contain a detailed discussion to enable an in-depth review of the specific technical issues. Recommended length for modules / mission mode projects are given in brackets.

- Objective (1-1 page). Succinct but clear statement of the project and modules; an estimate of quality measures must be provided for the entire project as well as the modules.
- Background (3-10 pages): International and national work on similar problems.
- Plan summary (3-5 pages): Rationale for the approach, key ideas, relation to other work.
- Preliminary results (3-5 pages).
- Detailed work plan (10-30 pages).
- Interfaces (2-10 pages) : Technical Interfaces needed to integrate with other modules / systems
- Evaluation (5-15 pages). Appropriate metrics must be outlined.

Proposers are strongly encouraged to include in their proposals a process and resources to conduct metrics-based evaluation. For core engines, benchmarks must be provided. For module and data proposals, metrics must be given for primary and annotated data

Suggestive Terms:-

1. Technical Rationale. Analysis of critical challenges and proposed solutions.
2. Modules. A list of proposed modules or sub-tasks, including the principal investigator for each module.
3. Resources Required. Any resources, including linguistic data, required to accomplish the task that are not covered by another task in the same proposal. It may include suggestions for data of general interest whose acquisition the government could fund separately.
4. Outputs Required. Any outputs from other engines that are required to accomplish the task.
5. Work Plan. Details of how the work will proceed.
6. Outcome : Final finished product ready for delivery to masses.

Product : Final output matching specifications defined in the objective.

Annexure II

RFP: Machine Translation : (a) English to Indian Language and (b) Among Indian Languages

Department of Information Technology
Ministry of Communications and Information Technology
Government of India, New Delhi

1 Goal

The goal of this RFP is to solicit proposals to build Machine Translation System from English to Indian languages as well as amongst Indian languages. This Machine Translation system is expected to work as an independent system, along with existing Word Processor and also as a component in Information Retrieval. The system normally is rated as Simple Machine Translation, High Quality Machine Translation and General Machine Translation.

2 Approach

The architecture of Machine Translation would include modules such as Morph Analyser, Rule Based Lexicon, part-of-speech Taggers and Chunkers etc. These Taggers and Chunkers would be based on statistical machine learning techniques and Morph Analysers would be typical tools. Chunker mixture on rule based approach is suggested to account for structural similarities amongst Indian languages. The final output would be generated by generator in target language. The architecture should enable to include multiple target languages e.g. English to Hindi translation engine should be tunable to include other Indian languages as target language.

Rule Base: This contains rules for mapping structures of sentence from English to Indian languages. This database of patter-transformations from English to Indian languages in entrusted the job of making a surface-tree to surface-tree transformation, bypassing the task of getting deep tree of the sentence to be translated. This database of patterns formations from English to Indian languages is entrusted the job of making a surface-tree to surface-tree transformation, bypassing the task of getting a deep tree of the sentence to be translated. The data base of structural transformation rules from English to Indian languages forms the heart of the Anglabharti system.

Sense Disambiguator: This model is responsible for picking up the correct sense of each word in the source language. It should be of interest to note that sense disambiguation is done only for the source text.

Target text generator: These form the tail end of the system. Their function is to generate the translated output for the corresponding target languages. They take as input as intermediate form generated by other modules and convert it to presentable form as system output.

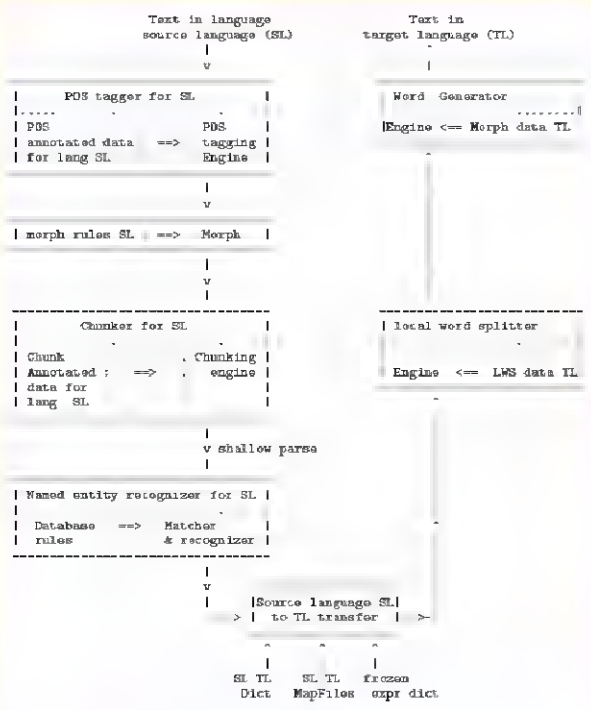
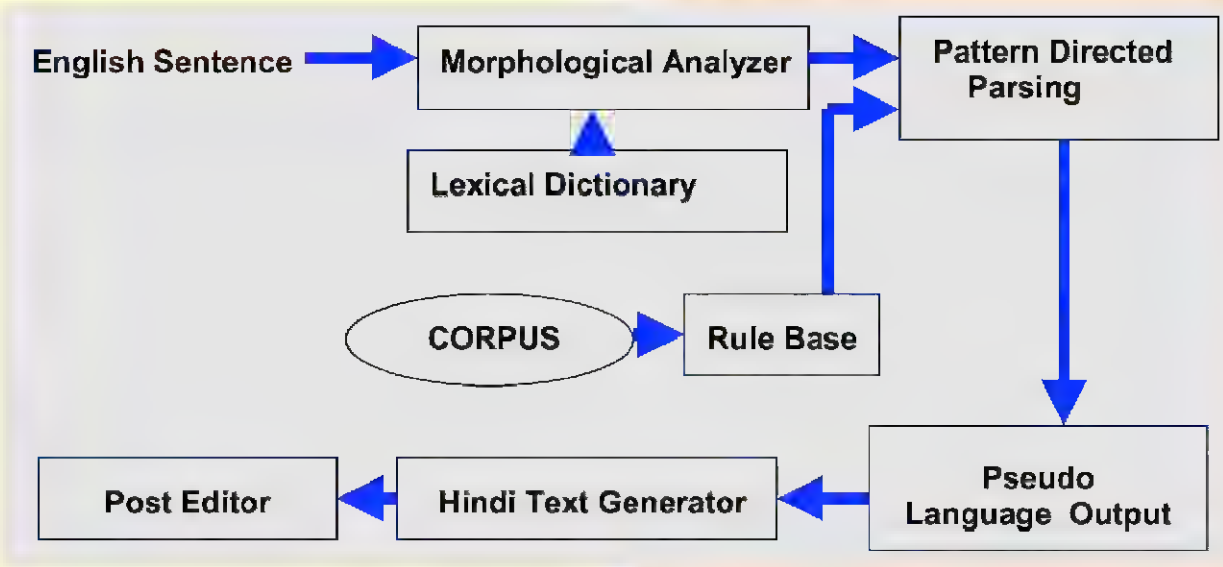


Fig. 1: Architecture of Simple MT System



3 System Architecture

The system architecture should include components which are tried and tested and which can be used in hosting of other applications e.g. for simple machine translation, the architecture could be as in Figure 1. The other alternative approach can be as in figure 2.

4 Components

4.1 Transliteration

A module which can perform transliteration among Indian languages, including Urdu, needs to be developed. Transliteration allows a word or words to be rendered in the script of the reader. For example, if a person who know Hindi reads Bangla text in Devanagari, he can still understand some parts of the meaning.

It can be seen that even when no other linguistic resources for translation are available among Indian languages, transliteration can still allow a reader to try to read and understand. Indian languages share a large number of lexical items, and simply by a change in the script the reader can understand quite a few things.

When linguistic resources are available, transliteration can still be used to render words which are not handled by the MT system to be rendered in the user's script.

4.2 POS Tagging Engine

The tagging engine should be independent of the language. Hidden Markov models (HMMs) would be used for building tagging engine. Robust and efficient implementation of HMMs is, therefore, required. The delivered engine should have appropriate smoothing functions appropriate for part-of-speech (POS) tagging. The trainability of the engine will be evaluated based on speed, memory besides precision.

Besides the engine based on HMMs, other proposals will also be considered.

4.3 Chunking Engine

Chunking engine will be the same as the POS tagging engine and will be based on HMMs. The only difference would be in the smoothing techniques adopted.

4.4 Morphological Analyzers &; Generators

Morphological analyzers and generators are required for each language. They should have a coverage of 85% of the language to begin with. Later new benchmarks would be defined going upto 95%.

4.5 Bilingual (Multi-lingual) dictionary

Two bi-directional bilingual dictionaries between Hindi and the concerned IL are needed (one in each direction).

Such a dictionary would actually be multi-lingual. However an agency working on language L would need to work, as if, it is preparing a bilingual dictionary only.

A reference dictionary template would be provided. Words and their major senses

would already be available in the source language. Such a reference dictionary would help to build a multilingual dictionary. The size of the dictionary would be 25,000 root words, and would be built in stages.

4.6 Transfer Engine

This component substitutes words using the bilingual dictionary as well as "simple" syntactic transformations. The transformation component could be specific to a language pair, however, for it to be practical it is likely to be specific to a pair of language groups. Examples of language groups are Dravidian languages (Tamil, Telugu, Kannada, Malayalam, etc), languages in the northern and western belt (Hindi, Marathi, Gujarati, Punjabi, etc.) and Eastern languages (Bengali, Assamese, Oriya, etc) which share greater com-monalities within their group

4.7 Annotated Corpora Preparation and Validation

For training the engines for the purpose of POS tagging and chunking, annotated corpora needs to be prepared. This ought to be a bootstrapping process in which after a small amount (say, 20,000 words) of initial corpus is tagged manually, the engine is trained on it yielding a preliminary tagger.

The tagger is now used to actually tag the texts, which are then manually corrected by human taggers. The new data so obtained is used to train the tagger virtually every night, and then repeating the process.

This is a rapid way to create tagged corpus as well as a tagger. In other words, data as well as technology components are created both at the same time. For monitoring the development, it is good to also create testing data against which benchmarks are run (every night).

4.8 Evaluation of MT system

Evaluation mechanism needs to be setup for end users for individual sentences as well as text. This should allow human or subjective evaluation, in the first stage, of comprehensibility and quality of the translation.

In the second stage, evaluation would need to be done on speed, usability, installation ease, etc. of the system.

5 Application Procedure

Interested educational institutions, R&D institutions or companies may submit proposal for developing the entire system or a component thereof, specifying Indian languages for which the work would be done.

In case of engines, which are independent of language, an Indian language must still be specified using which the monitoring and testing of the engine would be done. It must be clearly specified whether the language data would be prepared by the proposing agency or data existing with other groups would be used. In case of the latter, a letter in support of the availability of data for evaluation purposes must be included along with the proposal.

The interested institutions may submit an expression of interest, within 2 weeks of the issue of RFP. The expression of interest should describe component(s) that an institution is bidding for. Technical details should be given of the component(s), and, full details of the approach, in case, it is different from the one given here. Description of the capability of the investigators and past work done, must also be provided. Financial bid is not needed at this stage.

Based on the expressions of interest, a meeting would be called of the interested institutions. After that a detailed proposal with financial details would be required.

The final proposal would have to give financial details, as well as the full details of the planned approach to be used in building the component (s), comparison with state of the art or similar work done elsewhere, evaluation criteria to be used for continuous monitoring of progress, capability of the agency and that of the principal and other investigators for undertaking the work. See elsewhere for technical details required for the final proposal.

The final proposal would have to be submitted in a standard format which can be downloaded from the TDIL website http://tdil.mit.gov.in

Annexure III

RFP: Cross Lingual Information Retrieval with English to Indian Language Simple Machine Translation Capability

Department of Information Technology
Ministry of Communications and Information Technology
Government of India, New Delhi

1 Goal

The aim of this RFP is to solicit proposals to make the web more accessible to Indians by building technology that allows the user to give queries in an Indian language, and retrieve documents in other languages (principally English) as well, by converting/translating them to the language in which the query was issued. However, development of the high quality MT capability is not a part of this RFP.

2 Approach

The system will use a combination of information retrieval, keyword translation and other related technologies, to provide a cross-lingual information retrieval system between English and an Indian language (referred to as 'IL' in the rest of this document). The initial domain would primarily be of business and popular science.

The system will accept user queries in an Indian Language, translate them into English as well, and run the queries against a web search engine such as Google.

The quality of retrieval will depend on the domain of the query and the documents. The system will be tuned for documents about business and popular science, and will give reasonable quality of retrieval for these domains. Its output may be less accurate for general queries.

To show retrieval results, document summarization technology would also be developed.

3 System Architecture

System architecture (Fig. 1) shows data flow between components. The components included here are based on tried and tested methods. The components are usable in a host of other applications, and may also be needed in the other RFPs. The architecture shown is a schematic, and it may used as given or its variant. Another approach may be based on UNL (Universal Networking Language), that supports 15 world-languages basically for translation of technical material. This uses Universal Word Dictionary (UWD) and En-Converter (EC) and De-Converter (DC) for user language. For 3-4 Indian languages, EC, DC & UWD may be developed. Some work has been done for Hindi. To begin with, we may focus on Hindi, Tamil and Bangla.

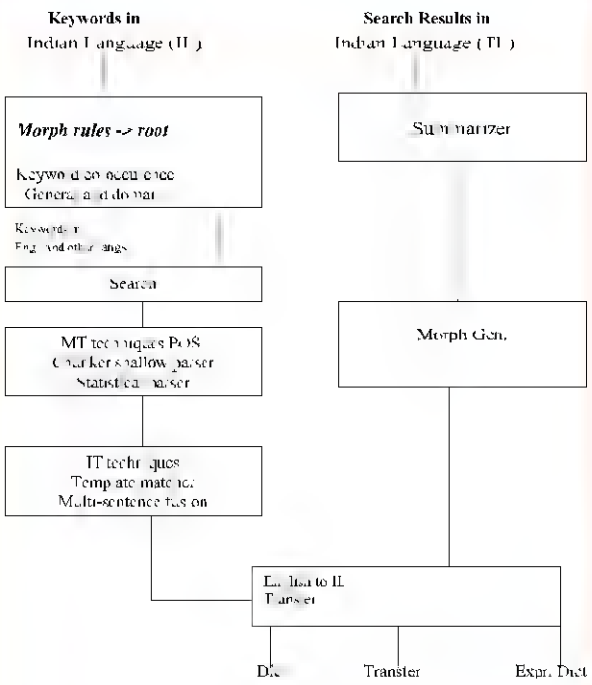


Fig. 1 : Architecture of CLIR

4 Components

4.1 Search mapping

Searches in ILs will be enabled by keyword translation. Good-quality morphological analyzers in the ILs would be required. Accuracy and precision exceeding 95% against a standard

corpus (such as the MCIT / CIIL corpora) would be necessary. A good general-purpose bilingual dictionary with 50K root words (multiple senses) is needed for keyword based searches. For example, the TransLexGram project provides a basis set of 15K sentences for some IL's. However, sentential co-occurrence frequencies of search keywords in the target IL are needed for disambiguation of searches initiated by multiple keywords. Supplementary dictionaries of 20K words each in the business and popular science domain would also be required.

Search results are retrieved using a standard search engine with the mapped keywords; google may be used as a reference.

4.2 English to Indian language (E-I) Dictionaries

As already referred to in the keyword search mapping section above, a 50,000 general purpose root-word English to Indian language dictionary needs to be developed. In addition, a 20K root word dictionary for each of the specific domains of business and popular science also needs to be developed on a frequency basis. Each sense in these bilingual dictionaries will be illustrated by an English sentence (derived from journals and magazines) followed by a manual translation in the IL. Sentences must be tagged with POS information, chunks, and parses in both English and ILs. While the sample sentences may be multi-clausal, the highlighted sense must occur in the main clause.

4.3 Named Entity Recognizers

Named entity recognizers must be developed for the two domains, namely popular science and business. The Penn treebank can be used for the business section, and Scientific American articles for popular science. The recall and precision of the NER's must exceed 85% on a representative cross section.

4.4 Morphological Analyzer in ILs

Morphological analyzers are required for each IL. They should have coverage of 85% of the language in the first phase (1 year), with a final performance of 95% against a standard corpus of about a million words.

4.5 Summarization Engine

Summarization engine should be able to take language data and components such as morph analyzers, POS taggers, thesaurus, etc. for a language, and become a summarizer for the language.

A summarization engine would be judged based on the quality of summary produced by it. Two engines would be compared by giving them the same resources for a language, and evaluating their performance on the same texts in the same language.

4.6 Evaluation of the system

Appropriate metrics must be outlined. Proposers are strongly encouraged to include in their proposals a process and resources to conduct metrics-based evaluation. For core engines, benchmarks must be provided. For module and data proposals, metrics must be given for primary and annotated data.

In the second stage, evaluation will be done on speed, usability, installation ease, etc. of the system.

5 Application Procedure

Interested educational institutions, R&D institutions or companies may submit proposal for developing the entire system or a component thereof, specifying Indian languages for which the work would be done.

In case of engines, which are independent of language, an Indian language must still be specified using which the monitoring and testing of the engine would be done. It must be clearly specified whether the language data would be prepared by the proposing agency or data existing with other groups would be used. In case of the latter, a letter in support of the availability of data for evaluation purposes must be included along with the proposal.

The interested institutions may submit an expression of interest, within 2 weeks of the issue of RFP. The expression of interest should describe component(s) that an institution is bidding for. Technical details should be given of the component(s), and, full details of the approach, in case, it is different from the one given here. Description of the capability of the investigators and past work done, must also be provided. Financial bid is not needed at this stage.

Based on the expressions of interest, a meeting would be called of the interested institutions. After

that a detailed proposal with financial details would be required.

The final proposal would have to give financial details, as well as the full details of the planned approach to be used in building the component (s), comparison with state of the art or similar work done elsewhere within the country and abroad, evaluation criteria to be used for continuous monitoring of progress, capability of the agency and that of the principal and other investigators for undertaking the work.

The final proposal would have to be submitted in a standard format which can be downloaded from TDIL website: http://tdil.mit.gov.in.

Annexure IV

RFP: Optical Character Recognition (Handwriting Recognition)

Department of Information Technology
Ministry of Communications and Information Technology
Government of India, New Delhi

1 Goal

The goal of this project is to build robust character recognition systems for content creation for applications such as digital libraries.

1. OCRs/OHR to convert images to text; to handle
   - Page layouts commonly seen in books, manuscripts and news paper articles;
   - Work efficiently for Indian Languages for all common (non-fancy) fonts and styles;
   - Provide highly accurate text for printed books and generate output in formats like html, rtf, pdf etc.; and
   - Mechanism to learn and adapt over time to improve the performance.

2 Approach and Architecture

OCR/OHR share many common modules irrespective of the language and script. OCR engines will be built with a common, language-independent structure. Language specific modules will be plugged in at the appropriate levels.

The problem will have four broad stages: (a) Page layout analysis and synthesis (b) Word segmentation to primary visual components (character or sub-character level) (c) Recognition of visual components and (d) Post-processor for improving the accuracy. Engines for each of these tasks will be developed and demonstrated on a spectrum of pages. Regular evaluations will be carried out. The code, resources, and results will be shared across various groups.

Modularisation: The development will take place in a modularised manner with well-defined interfaces across modules. This will enable groups to work independently on the modules and will facilitate easy integration and building of the system at any stage.

Performance Parameters: Appropriate evaluation measures and data sets will be defined. This will allow the performance to be evaluated at module-level. The data set for evaluation will be selected from the Digital Library of India content by an expert team. Ground-truth data will be provided for automatic performance evaluation.

Evaluation Process: Performance of the modules will be evaluated on ground-truth data sets, which is also generated as part of this project from Digital Library pages.

3 Components

3.1 Page Layout Analysis Engine

Page layout analysis module will accept images in standard input formats like TIFF, PNG, JPG, GIF, BMP, etc. Books will be assumed to be in the standard format used by the Digital Library of India.

Page layout module analyzes the images and generate an output description with text and graphics blocks well represented, textual blocks further divided into individual words and bounding boxes of the word boundaries generated. Details of the words like size, style, script, alignment of the blocks; layout of the page, etc. will be retained in a common structure. Output format will be finalized to suit the standard open architecture. In addition modules for synthesis of the html/rtf pages retaining text information will be generated.

3.2 Visual Component Extraction Engine

Word-images are segmented to consistent visual component by this module. Segmentation algorithm accepts the complete document image layout information, and the word coordinates. It outputs a set of visually distinguishable components consistently for all fonts/sizes and styles for a language. Implementation will have to be robust and consistent. Algorithms will have to

be reasonably robust to various popular degradations seen in the images.

In addition, this module provides a table-lookup for relating an Akshara to the consistent sequence of components it corresponds to. Number of visible components will have to be small. Algorithm for segmentation may depend on the language.

3.3 Visual Component Recognizer Engine

This module provide support for efficient feature extraction and classification. In its offline phase, it can train the system with labelled examples. During the test phase, it provides n-best matches in a ranked form for each of the components, with confidence scores like posterior probabilities.

This engine will have to be fast and highly accurate. Also there should be provision for adapting to a specific collection of books if needed, perhaps with the help of additional training data. This module also integrates the classification results of various components and must provide a textual string in Unicode / ISCII and romanized version in INSROT.

3.4 Post Processor Engine

Post processor engine for OCR is different from the post processors available for word processor. It needs to verify whether a given word is present in the language. This may use large corpus and language modules. This module may store the dictionary and resources in multiple files based of the frequency of the use to speedup the process.

This module supports correction to the recognised text (or selection from the possible strings) in presence of errors in classification like substitution of components / characters, additional components / characters, missing components / characters.

Post processor module is not expected to perform sentence level processing.

3.5 Training and Testing Data

Annotated data is required to train and test the engines and the modules of the system. It is planned to generate large quantity of annotated ground-truth data from multiple real-life documents.

Annotation Tool: A tool, which accepts input images and semi-automatically segments, labels the blocks; words and components are needed for the rapid generation of the test data. This tool outputs the ground-truth data in an open user-friendly format.

Generation of Data: The annotation tool can generate training and testing data on pages in multiple scripts, layouts, styles, etc. The tool may work in a semi-automatic or fully automatic mode. Selection of pages for annotation, their quality, and format will be performed with inputs from experts in this area.

Early versions of the engines can be plugged into the tool for improving the semiautomatic labelling using a bootstrap strategy.

4 Application Procedure

Interested educational institutions, R&D institutions or companies may submit proposal for developing the entire system or a component thereof, specifying Indian languages for which the work would be done.

In case of engines, which are independent of language, an Indian language must still be specified using which the monitoring and testing of the engine would be done. It must be clearly specified whether the language data would be prepared by the proposing agency or data existing with other groups would be used. In case of the latter, a letter in support of the availability of data for evaluation purposes must be included along with the proposal.

The interested institutions may submit an expression of interest, within 2 weeks of the issue of RFP. The expression of interest should describe component(s) that an institution is bidding for. Technical details should be given of the component(s), and, full details of the approach, in case, it is different from the one given here. Description of the capability of the investigators and past work done, must also be provided. Financial bid is not needed at this stage.

Based on the expressions of interest, a meeting would be called of the interested institutions. After that a detailed proposal with financial details would be required.

The final proposal would have to give financial details, as well as the full details of the planned approach to be used in building the component (s), comparison with state of the art or similar work done elsewhere, evaluation criteria to be used for continuous monitoring of progress, capability of the agency and that of the principal and other investigators for undertaking the work. See elsewhere for technical details required for the final proposal.

The final proposal would have to be submitted in a standard format which can be downloaded from the TDIL website http://tdil.mit.gov.in

Annexure V

Gist of Tools available from the institutions funded by TDIL in formulating the components of BIPK

LANGUAGE	TOOLS											
	FONTS		OPEN OFFICE		WORD PROCESSOR		DICTIONARY		E-MAIL CLIENT		BROWSER	
LANGUAGE	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER
PUNJABI	AVBL.	CDAC			AVBL	RC	AVBL.	RC				
MALYALAM	AVBL.	CDAC			AVBL	RC	AVBL.	RC	AVBL	RC		
TELUGU	AVBL.	CDAC			AVBL	RC	AVBL.	RC				
ORIYA	AVBL.	CDAC			AVBL	RC	AVBL.	RC	AVBL	RC		
BENGALI	AVBL.	CDAC										
KANADA	AVBL.	CDAC	?	CDACB	AVBL	RC	AVBL.	RC				
ASSAMESE	AVBL.	CDAC					AVBL.	RC				
MARATHI	AVBL.	CDAC	?	CDACB								
URDU	AVBL.	CDAC			AVBL	RC						
GUJARATI	AVBL.	CDAC	?	CDACB								

Annexure VI

Gist of Tools available from the institutions funded by TDIL in formulating the components of BIPK

LANGUAGE	TOOLS											
	MACHINE TRANSLATION		OCR		TEXT TO SPEECH		TRANSLI-TERATION TO ROMAN		TEXT CORPORA		SPEECH CORPORA	
LANGUAGE	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER
PUNJABI			AVBL	RC			AVBL*	RC				
MALYALAM			AVBL	RC	AVBL	RC						
TELUGU			AVBL	RC	AVBL	RC						
ORIYA	AVBL	RC	AVBL	RC	AVBL	RC						
BENGALI	AVBL	CDACK	AVBL	RC	AVBL	CDACK			AVBL	RC	AVBL	CDACK
KANADA	AVBL	RC			AVBL	RC						
ASSAMESE			AVBL	RC					AVBL	RC		
MARATHI				DEVA-NAGARI OCR**					AVBL	RC		
URDU												
GUJARATI												

* Gurumukhi - Roman

Gurumukhi - Shahmukhi

** Devanagari OCR may be used

Gist of Tools available from the institutions funded by TDIL in formulating the components of BIPK

LANGUAGE	TOOLS											
	SPELL-CHECKER		MORPHOLOGICAL ANALYSER		THESAURUS		ESTIMATION PACKAGE		WEB CONTENT		WORDNET	
LANGUAGE	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER	AVAILABILITY	DEVELOPER
PUNJABI	AVBL	RC										
MALYALAM	AVBL	RC					AVBL	RC	AVBL	RC		
TELUGU	AVBL	RC	AVBL	RC					AVBL	RC		
ORIYA	AVBL	RC	AVBL	RC								
BENGALI	AVBL	RC			ABVL	RC						
KANADA			AVBL	RC					AVBL	RC		
ASSAMESE	AVBL	RC										
MARATHI											AVBL	RC
URDU												
GUJARATI												

Annexure VIII

LIST OF RCs

LANGUAGE	INSTITUTIONS									
	TIET, PATIALA	CDAC TRIVANDRUM	UNVIERSITY OF HYDERABAD	UTKAL UNIVERSITY ORISSA	ISI, CALCUTTA	IISc, BANGALORE	IIT, GUWAHATI	IIT MUMBAI	CDAC PUNE	MS UNIVERSITY BARODA
PUNJABI	YES									
MALYALAM		YES								
TELUGU			YES							
ORIYA				YES						
BENGALI					YES					
KANADA						YES				
ASSAMESE							YES			
MARATHI								YES		
URDU									YES	
GUJARATI										YES

Annexure VIII

TECHNOLOGY DEVELOPMENT FOR INDIAN LANGAUGES (TDIL)
Committee Members

1.	**Shri Brijesh Kumar** Secretary DIT, MCIT, New Delhi	Chairman
2.	**Prof. N. Balakrishnan** Chairman Division of Information Science Indian Institute of Science, Bangalore	Member
3.	**Prof. S. Dhande** Director IIT, Kanpur	Member
4.	**Prof. S.V. Ramanan** The AU-KBC Research Centre Madras Institute of Technology Anna University, Chennai 600 044	Member
5.	**Prof. Rajeev Sangal** Director, IIIT Gowchibowli, Hyderabad 500 019	Member
6.	**Shri Manoj Annadurai** Shakti Technologies 2, Reddy Colony, Ramalinga Puram, Chennai	Member
7.	**Shri R. Chandrashekhar** Joint Secretary DIT, MCIT. New Delhi	Member
8.	**Shri Ajeer Vidya** Joint Secretary & Financial Advisor DIT, MCIT, New Delhi	Member
9.	**Ms. Swaran Lata** Director DIT, MCIT, New Delhi	Member
10.	**Dr. Hemant Darbari** Programme Coordinator, GIST C-DAC Pune University Campus, Pune 411 007	Member
11.	**Shri V N Shukla** Director (Spl. App.) C-DAC, Noida	Convener

3. National Roll Out Initiative Uniting People through Indian Language Technologies

Computer technology in India has both a developmental as well as a social role. In its developmental role, it is concerned with the designing and development of newer technologies for various applications. In its social role, it breaks the language barrier and bridges the gap between the various sections of the society through easier access to information using their respective mother tongues or local languages. Language here has a major role to play and, therefore, language computing becomes central to the exchange of information across speakers of various languages. India is a multilingual country with as many as 22 scheduled languages and only about 5% of the population is able to understand English.

Keeping this in mind, Ministry of Communications and Information Technology launched a major initiative in the year 2005 the area of e-governance, to enable more reliable and efficient services to our citizen's requirements from the government in tasks ranging from land / citizen information to obtaining passport and managing tax payments. Through this Language Technology mission, a major initiative has been taken to aggregate Indian Language fonts and software tools from various public / private players, incorporate them into user friendly tools and products and make them available free for public through CDs and web downloads. IT ministry's initiative of Technology Development for Indian Languages (TDIL) has been instrumental in generating wide interest in developing technology and resources relevant to the use of Indian languages in ICT (Information and Communication Technology). Through a number of resource centres and funded projects, it has created rich human resources, linguistic resources, software tools, etc. across the country in many languages.

Current releases

C-DAC, GIST, Pune under the leadership of TDIL, DIT has released the CD's for ten constitutionally recognized Indian languages viz. Tamil, Hindi, Telugu, Marathi, Urdu, Punjabi, Oriya, Kannada, Assamese and Malayalam for free mass usage . Gujarati and Bengali are in pipeline. The process of consolidation for other languages as well as release of second version of CD for some of the languages is currently underway.

Since the CD is targeted towards common man, it contains tools for common man, productivity enhancement tools & beta tools for getting more feedback from user for future research.

Broadly the contents of each language CD are as follows:

1. True Type fonts with keyboard driver
2. Multi-font keyboard engine for True Type fonts
3. Unicode compliant Open Type fonts
4. Unicode compliant keyboard driver.
5. Generic font's code and storage code converter
6. Localized version of BharateeyaOO (Spread Sheet, Presentations, Word processing & drawing tools), Fire fox Browser, Thunderbird email client, GAIM (Multi protocol messenger)
7. Spellchecker
8. Bilingual Dictionary
9. Decorative fonts design tool
10. Transliteration Tool
11. Language Tutor
12. Text to Speech
13. Database sorting tool
14. Microsoft word tools
15. Microsoft Excel tools
16. Type Assistant
17. Content management system
18. Typing tutor
19. Games / Puzzles
20. Library Management System
21. Seamless email send / receive utility
22. Text Editor
23. OCR
24. WorldNet
25 Morphological Analyzer and Generator
26. Text To Speech Systems

2.0 Process / life cycle

The tools built by resource center along with contributions from other academicia and private Indian language developers have been consolidated into the "free software tools and fonts". These are available in the form of a free CD as well as free download on the internet. On

registering online free home delivery of CD is done anywhere in India.

The contributors were identified by releasing press advertisements from time to time. A performa for submitting contributions was made available. On submission these possible contributors were invited for technical deliberations.

A technical team evaluated the proposals as well as the software and negotiated finances with the possible contributors, for finalizing the list of software. Technical constraints such as size of software, usability, novelty / availability of other similar software, impact of the technology, etc. were also considered.

The software submitted by the possible contributors was vetted for basic functionality and use. The resulting test reports were submitted back for possible corrections and updates. In several cases the contributors worked very hard and updated softwares in extremely short periods of time to cater to the needs of quality and usability.

3.0 Localization an Overview

Localization of Bharateeya Open office in all 22 Scheduled Indian languages

Localization is the process of adapting a product or service to a particular language, culture, and gives the desired local "look-and-feel." A successfully localized service or product is one that appears to have been developed within the local culture. Before we discuss the above process in detail, we must talk about the languages and the importance of availability of all data including the IT related data in all languages existing in the world today so that progress reaches equally to all sections of society.

Language is a way of communication between two people or between a group. Ethnologue lists 6,912 living languages in the world today though it may never be determined exactly as there may be many more civilizations which are yet to be listed. As a large and linguistically diverse country, India is listed to have around 427 languages languages out of which 22 languages are officially recognized. IT field is developing at a massive speed but all the data is available mostly in English

To show the necessity of availability of localized versions of computer material, we are listing below our 22 official languages, with the places they are spoken along with the number of speakers (wherever data was available):

1.	Assamese	15 million	Assam
2.	Bengali	67 million	Andaman & Nicobar Islands, Tripura, West Bengal,
3.	Bodo		Assam
4.	Dogri		Jammu and Kashmir
5.	Gujarati	43 million	Dadra and Nagar Haveli, Daman and Diu, Gujarat
6.	Hindi	180 million	Andaman and Nicobar Islands Maharashtra Arunachal Pradesh, Bihar, Chandigarh, Chhattisgarh, the national capital territory of Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttarakhand
7.	Kannada	35 million	Karnataka
8.	Kashmiri	56,693	Jammu and Kashmir
9.	Konkani	1,760,607	Goa, Karnataka, Maharashtra
10.	Maithili	22 million	Bihar
11.	Malayalam	34 million	Kerala, Andaman and Nicobar Islands, Lakshadweep
12	Manipuri	1,270,216	Manipur
13.	Marathi	65 million	Dadra & Nagar Haveli, Daman and Diu, Goa, Maharashtra
14	Nepali	2,076,645	Sikkim, West Bengal
15.	Oriya	30 million	Orissa
16.	Punjabi	26 million	Chandigarh, Delhi, Haryana, Punjab
17.	Sanskrit	49,736	
18	Santhali		Santhal tribals of the Chota Nagpur Plateau (comprising the states of Bihar, Chattisgarh, Jharkhand, Orissa)
19.	Sindhi	2,122,848	
20.	Tamil	66 million	Tamil Nadu, Andaman & Nicobar Islands, Kerala, Puducherry
21.	Telugu	70 million	Andaman & Nicobar Islands, Andhra Pradesh
22	Urdu	46 million	Andhra Pradesh, Delhi, Jammu and Kashmir, Uttar Pradesh

In the above list, there must be a negligible percentage of people conversant in their local language as well as English. Hence, if the modern IT technology, wherein all data is available on a finger-tip, needs to reach the masses, the process of localizing all the data available is very important.

Realizing the need for localization, CDAC, GIST was given the assignment of releasing localized versions of Bharateeya OO, Fire fox, Thunderbird and Pidgin softwares in all the above languages. This was taken as a challenge, and the process of looking for experts / linguists for these languages began.

Challenges faced during Localization

Looking at the list above, you can see that, while there was lot of speakers for some popular languages, some languages have very less number of speakers which could not even be listed. Even amongst those, people who were computer savvy, conversant in English and also the local language were very less. While there were lot of literary experts in various languages but in the computer field, very few were available. Moreover, localization was meant for the common men, who were not supposed to be fluent in English and hence we were actually required to create a whole new set of words especially for computer usage which have to very user-friendly. Hence, keeping this in mind, we had to search for competent people.

The material provided to us for localizing purposes were a set of 60000 basic computer strings in English and 2 lakh strings for advanced users, which needed to be localized (not translated) in all languages. Moreover, these strings were not complete sentences, wherein you can find tenses, verbs or clauses. They were computer commands, which we use in English. Even the basic words like file, folder, directory, document etc. were not enlisted in any dictionary available. While in some languages, it was transliterated and retained as they were in English, linguists / experts of certain languages set about the task of creating words in their own language for every word in IT terminology. Then, there were words like delimiters, add-ons, plug-ins etc. which had become words of common usage in English, but for other languages, finding equivalent words was a tough job. This was a mammoth job, involving lots of discussions and mutual agreement. These localized words were just the consolidated opinion of some experts, but if these words were to reach the common man, dictionaries were necessary. Simultaneously, a glossary was also developed for all languages for this purpose.

Within every language, dialects were different in the cities of the same state. We also had to take this into consideration, while making the strings. For instance, Hindi spoken in different parts of our northern belt varies a lot from city to city. Hence, this was a continuous experimentation process to find words commonly acceptable to all.

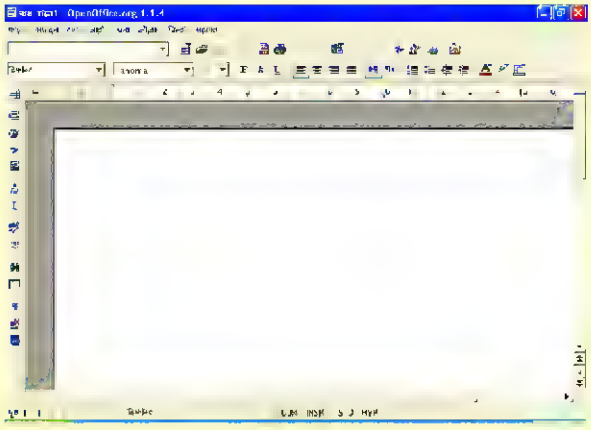
Pune, being the hub of education, where lot of students from all over India come to pursue their higher studies and do PhDs. We were fortunate to find some local experts for Sindhi, Urdu, Assamese, Oriya , Marathi, Bengali etc. and through their help, experts already working in this field were enlisted from local areas. Various universities, wherein linguistics was taught as a subject were approached, and advertisement inviting applications was released in various national and local newspapers. We presently have expert team of 70 free-lancers working in various languages from different parts of India.

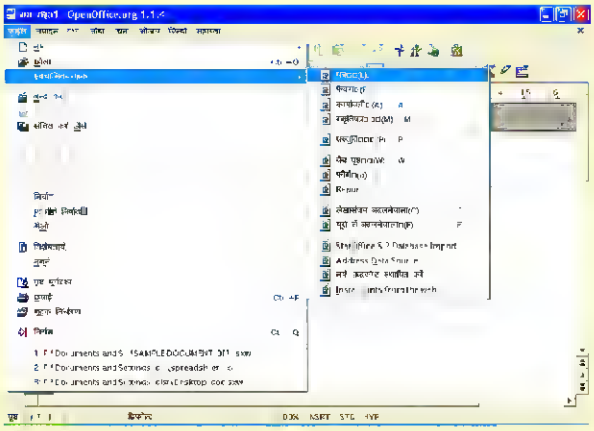
The real challenge was to find experts in languages like Dogri, Maithili, Santhali, Bodo and Sanskrit. Lots of literary work has been done in these languages but no terminology or even basic dictionaries were readily available in these languages. We could find some experts, but the speed of work was slow as it was a completely new area of work for them. But now basic translations are over in all the above languages except Sanskrit and validation is going on now. Sanskrit localization has also been taken up by JNU at Delhi.

4.0 List of Contents / Contributions

The free software tools and fonts CD consists of various tools and technologies classified as under

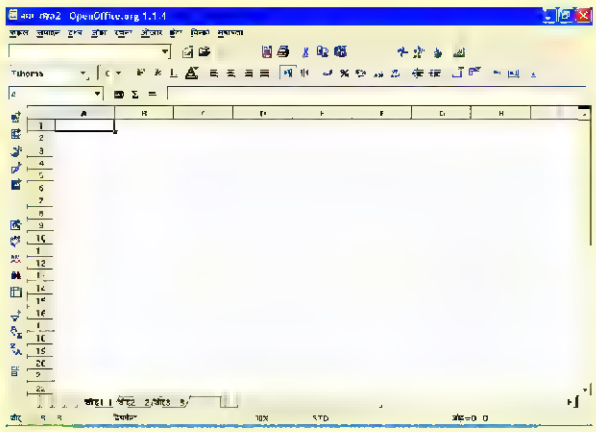
4.1 Basic Information Processing Kit
4.2 Productivity enhancing tools
4.3 Beta tools for power user

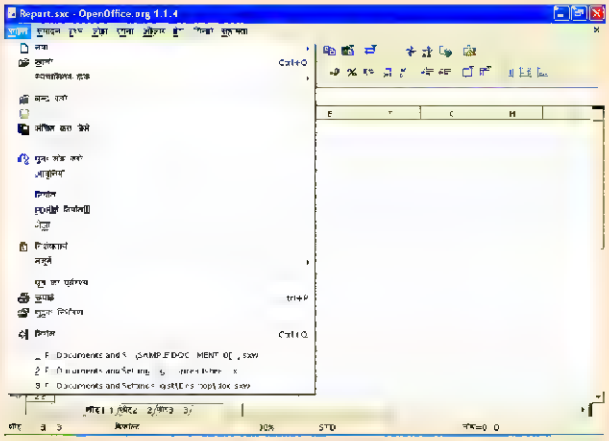




Bharateeya Open Office Calc

Calc is a spreadsheet program. A spreadsheet program enables us to store and manage our numeric data, and perform calculations in it. Calc has a wide array of tools and features that enable us to perform complex calculations with just a few mouse-clicks. It can also open and save other spreadsheet application files such as MS-Excel.



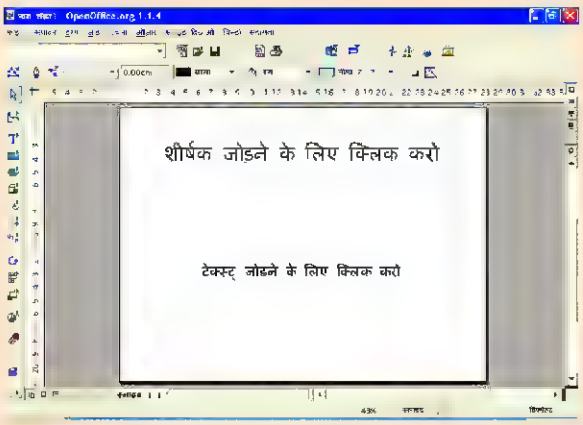


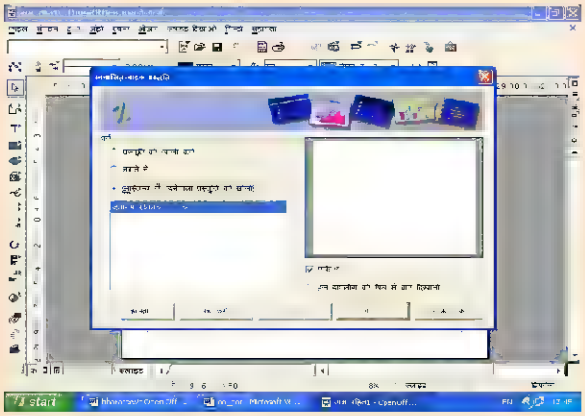
Bharateeya Office Impress

It is a presentation application that provides us with many tools for creating presentations. Presentation software helps us to create dynamic slides and present them to the audiences through a slide show. Useful presentation software that is included in the BharateeyaOO package is Impress. It is a very efficient and potent program that offers a number of features that enable us to create and modify effective and appealing presentations and run them in a slide show.

The tools include, guides for positioning objects, automatic snapping of the object to a freely definable grid or to each other, and scaling and cross-fading and many other effects.

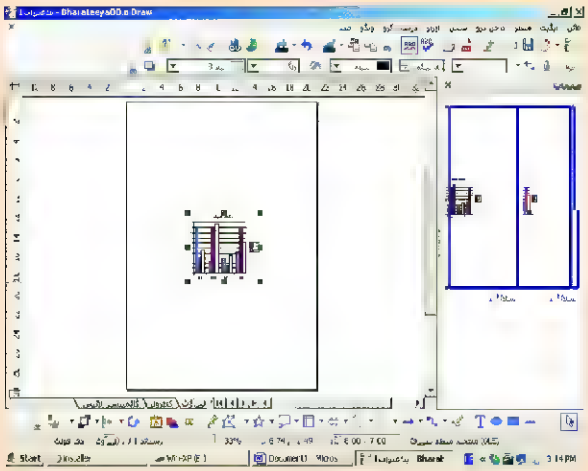
There are also other features in BharateeyaOO - Impress, which can help us create our presentations easily and quickly. We can also export our presentation to be published on the Internet. All the necessary conversions are done automatically. The presentation you export can be viewed with any modern browser.





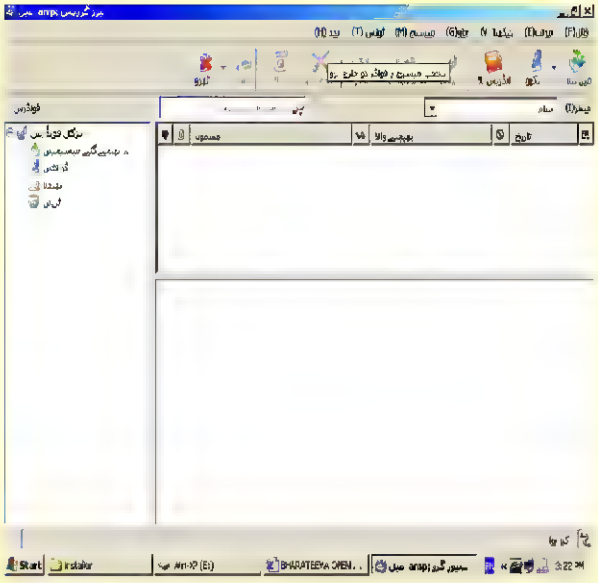
Bharateeya Open Office Draw

Draw is an object-oriented vector graphic drawing program. The object can be lines, rectangle, 3D cylinders or other polygons can be drawn. All objects already have set properties, such as size, color of the surface, color of contours, linked files, associated actions when clicked and much more. All of the properties can be modified at any time. The Bharateeya Draw has all UI in Indian languages. Image below shows Draw having all the menus localized in Urdu.



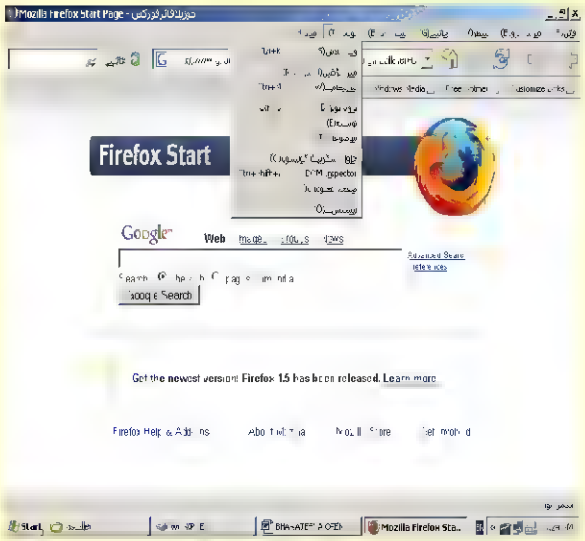
Thunderbird

Thunderbird is a free, open-source and cross platform mail client for most OS including, Windows, Linux and Macintosh. It is a complete email application, which is simple to use, powerful and customizable. It easily exports your existing email accounts and messages. It is similar to competing products like Outlook Express, but with additional features such as junk mail classification.



Firefox browser

The popular FireFox Browser has been provided with all the menus, status bars, error messages, user prompts, etc. in Indian Languages. The image below shows the Firefox browser, having all menus and various options, all localized in Urdu language.



5.0 Impact of these Software tools and technologies

Till 22/10/2007 approximately 293238, free software tools and Fonts CDs have been distributed to the masses. Apart from Online registrations and shipment of CDs, approximately

The CD has an Indian language Graphical User interface to facilitate use and instructions related to installation of the various software. The Graphics of CD and the CD cover have been designed keeping in view the cultural aspects associated with the language or the area where the language is most widely spoken or used.

4.1 Basic Information processing Kit (BIPK) for the masses.

It consists of Indian language tools and technologies which are required by the majority of users, including office automation, fonts and data entry and content creation tools.

The BIPK consists of highly calligraphic True Type and UNICODE compliant Open Type fonts of various languages, keyboard drivers and layouts for inputting.

Akruti Multi-font keyboard engine basically will be able to support typing and creating documents using any of the fonts supported by the tool. Basically there are two types of fonts which are popular for desktop usage. These are True type and Open Type fonts. True type fonts are basically vendor dependent. This means that the code used for storing the characters is designed by individual organization. To support typing using the said font, we need to have the driver which is developed by that organization only. Tomorrow if someone wants to edit the document already created, he will also need to have the same driver else it will not be possible for him to edit that document. As opposed to True Type, there are something called Open Type or Unicode enabled fonts. Unicode consortium specifies the codes for each character for all the languages worldwide using a unique code for each character. Therefore it becomes very simple for editing the documents created using Unicode fonts.

It also includes Localized Indian version of Open Office which includes word processor (writer). Spread sheet (calc), presentation tool (impress), drawing and math tool. For more information please see point 4.4

The Firefox browser, thunderbird email client and GAIM multiprotocol messenger have also been localized into Indian languages.

Keyboard and Language Learning tools have been included in the CD for proliferation of use of computers in Indian languages.

4.2 Productivity enhancing tools

Apart from the BIPK, the CD also consists of productivity enhancing tools such as Dictionaries, spellcheckers, legacy code converters, etc. The language learning, multimedia and keyboard learning tutor software are also included. We feel that the tools like Dictionaries are very helpful at various places, may it be office work or even for that matter building vocabulary of your child. Officials often create the documents in different languages and sometimes because of various reasons we need help of spell-checkers to see if the created document does not have any wrongly spelled words. Document converters have their own importance.

Today there are various companies involved in Indian language computing. Some of them are Akruti, Modular InfoTech, C-DAC, IMRC, C. K. Technologies, Softview, etc. These companies have their fonts in the market and used by lot of people especially in the areas of Data Entry, desktop publishing, media and others. As seen in the Multi-font keyboard engine explanation, people should have the drivers available with them to use the particular font in their documents. It may happen that the document which is created by one person needs to be edited / changed by other. In this case there is a dependency that the driver which was used for creating the document is also needed for editing the document. Also if someone wants to use another font from different vendor (may be because that font is better than the current font) then there should be a utility which can convert into the other font. Otherwise he will have to retype the entire document again. This is not that someone would really want because the amount of time and money and other resources required would be huge.

This tool converts various 8-bit True Type fonts encoding to and from ISCII / UNICODE. This is especially used when the data is created using font encoding.

4.3 Beta Tools (to update)

It takes lot of research and other efforts to develop or build tools such as optical character recognizer for Indian languages. Such tools should be used and evaluated by as many people as possible so as to make it more versatile. We generally call these products as Beta Products. Beta tools and research material such as Text to speech systems, Optical character recognition systems, morphological analyzers and generators etc. have also been included so that these are used by the experts, common people, Researchers, evaluate them and give the feedback to us so as to carry out further work and research to make these products better and useful.

4.4 Localized Open Source Software

Under this initiative C-DAC along with other open source partners has undertaken the task of localizing the following Free and Open Source Software:

- BOO which is an entire office suite and includes
  - Writer (word processor),
  - Calc (Spreadsheet),
  - Impress (presentation tool),
  - Draw (drawing tool)
- Thunderbird (E-mail client)
- Firefox (Browser)
- GAIM (Multiprotocol messenger)

The languages for which these are already available include: Assamese, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu and Urdu. These are available free and are data compliant with MS-Office products. Supports Linux as well as windows and the data generated is cross platform.

BHARATEEYA OPEN OFFICE SUITE

Open Office package is a very handy and useful set of application software that enables us to complete our day-to-day tasks quickly and efficiently. OO is an Office automation package, which contains a set of applications capable of performing a number of tasks like word-processing, spread sheets, etc. It is platform independent i.e. OO can work on Windows and Linux. Also being open source software, it is free software, unlike Proprietary Office suite

The BharateeyaOO suite mainly consists of a word processor application (Writer), a spreadsheet application (Calc), a presentation application (Impress), a drawing application (Draw). These applications have interfaces similar to those of the Common and widely used Office applications and those who are familiar with the other office packages such as, MS Office, Star Office and Lotus Smart Suite would find it easy to work with the BharateeyaOO package. The data generated is cross-compatible i.e. Existing Office data can be opened and changed using OO and vice versa. BharateeyaOO is UNICODE compliant and requires Indian language enabled OS. Indian Language input is possible using Open Type fonts (for languages supported in the OS) as well as True type fonts for others.

The BharateeyaOO suite, is the customized version of Open Office, with all the menus, status bars, error messages, user prompts, etc. localized for Indian languages. There is a Help menu in each of the BharateeyaOO applications, which contains help options in Indian Languages that you can use to get help about the various tools and commands of the application.

Bharateeya Open Office Writer

Writer is a free, open-source and cross platform (supports) word processor application, with all the menus localized. This is a word processor application, with all strings localized in Indian languages, which lets us create text documents, open existing MS-Word documents, Print, create Content, create WebPages, etc.
We can create personal letters, form letters, brochures, faxes and even professional manuals using this application. It provides a number of editing tools, which help us in editing and proofing our documents; and a number of formatting tools that help us in managing the appearance of our documents. It supports Windows, Linux as well as UNICODE. It can import as well as export files in .doc format as well as .html format. So Writer can be used as web-content creation tool.

0, 0, 0, downloads have happened from the website. Other media of distribution include magazines, pre-bundled software with OEM, etc.

Apart from individuals, several organizations such as Election Commision Of Indian, State Bank of Hyderabad, etc and e-governance initiatives and state and central government departments have migrated from using proprietary expensive software to using the free software tools available in CD, thereby saving the country's exchequer hundreds of crores of rupees annually.

5.1 CD distribution mechanism

The Department of Post has made special provisions and resources for delivering the free CDs. We are grateful to the Department of Post for ensuring that the impact of "free software tools and fonts" initiative reaches all corners of India.

5.2 Training on Indian Language Tools and Technologies

Inorder to further proliferate the use of the Indian language tools and technologies, training of various government departments such as Election Commission of India, Department of Information Technology, etc, Central Institute of Indian Languages (CIIL) , has been undertaken. Several PACE centers also offer low cost training to the masses on the tools and technologies offered under the initiative. These courses are targeted mostly at Indian language Office automation and content creation tools such as BharateeyaOO.o.

6.0 Major organization/contributors

We are grateful to all the contributors for offering their tools and technologies for a national cause. Several private developers have contributed their best commercially successful software tools and technologies in this national initiative.

We are also grateful to all the Academicians, University departments and Research Institutes who have contributed their years of research development work in Indian languages, thereby resulting in the success of the "free software tools and Fonts".

Several contributors have partnered with C-DAC in joint development, productisation and testing of the available tools and technologies. This includes Utkal university, University Hyderabad, Kannada Ganak Parishad / Govt. of Karnataka, etc.

Sr. No	Name of Contributor	Contact information
01	C-DAC	http www cdac in
02	Cyberscape M.l.med.a	http www cyberscapeind.a com
03	Chenna. Kav.ga.	http www snakt.off.ce com
04	Cadgraf D.gita.s Pvt Ltd	http www cadgraf com
05	IIT, G.wanat.	http www itg emet in
06	Kannada Ganak Parishath	http www kagapa org
07	IISc, Banga.ore	http www .sc emet .n
08	SoftV.ew Techno.og.es	http www softv.ew in
09	V.shwa Kannada Softech	http www v.shvakannada com
10	Thomson Cyber So. .t.ons	http www thomsoncyber com
11	Priya Informat.cs	http www priya.informat.cs com
12	IIT M..mba.	http www cf..t ..to ac in
13	Utka. Un.vers.ty	http /www utka.-un.versity org
14	P.n.ab Unives.ty, Patia.a	http www p.n.ab. .n.versity ac .n
15	Thapar Institute	http www t.et ac in
16	Mod.. ar Infotech	http www mod..ar-infotech com
17	NCPUL	http www .rd.co .nc.. n.c in
18	AU-KBC Research Centre	http www a..-kbc org
19	Learn-F.n Systems	
20	Pa.an.appa Brothers	http pa.brothers com
21	IIIT, Hyderabad	http www ..t net
22	Pro.og.x	http www pro.og.xsoft com
23	An..saraka	
24	CALTS	http www .onyd emet .n
25	Wassan	http www wassan org
26	Free software fo .ndat.on of ind.a	http www fsf org in
27	Utkarsh Magnet Techno.og.es	http www magnettechno.og.es com

The contributors have also been acknowledged in the respective CD and CD Covers.

7.0 Support and Feedback

We do not stop here just by developing the softwares, tools and giving to the common masses. We also consider that the help and support should be available to those people who are ready to learn and use these softwares and bring a joy amongst them by using these softwares in their own languages. For the same, we have a very strong team of dedicated people across different places of the country to provide help through various means like emails, telephones, faxes, letters, and even by providing classroom training wherever it is needed. People should not hesitate to contact us for any problem, however small it is.

Several Individuals and Organisation have expressed their satisfaction and appreciation of the free software tools and fonts CD initiative.

Some Useful Reference Links and acknowledgements

1. Technology Development of Indian Languages- http://tdil.mit.gov.in
2. Center for Development of Advanced Computing- www.cdac.in
3. Indian Languages Data Center- www.ildc.in/www.ildc.gov.in

Courtesy :

Ms. Swaran Lata
Director
TDIL Programme
e-mail: slata@mit.gov.in
Tel. : 011-24363525

Shri Mahesh D. Kulkarni
Chief Investigator
Email : mdk@cdac.in
Tel:020-25694000

4. हिंदी सॉफ्टवेयर उपकरण – भारत सरकार का महत्वपूर्ण कदम

कंप्यूटरी हिंदी के इतिहास में 20 जून 2005 का दिन ऐतिहासिक है। इस दिन भारत सरकार ने ऐसा कदम उठाया है कि जिसने हिंदी कंप्यूटरी के रास्ते में आने वाली सारी बाधाओं और समस्याओं का हल एक ही सीडी में उपलब्ध करा दिया। जिसका नाम था ''हिंदी सॉफ़्टवेयर उपकरण''। यह सीडी एकदम मुफ्त थी, जिसे आप निस्संकोच अपने मित्रों को वितरति भी कर सकते हैं। यह सीडी सूचना प्रौद्योगिकी में भारतीय भाषाओं के उपयोग को बढ़ावा देने के सूचना प्रौद्योगिकी विभाग और संचार एवं सूचना और प्रौद्योगिकी मंत्रालय के राष्ट्रीय प्रयास का एक कदम था।

''हिंदी सॉफ़्टवेयर उपकरण'' नामक इस सीडी को सरकार ने कितना महत्व दिया है यह इस बात से जाहिर हो जाता है कि इसका लोकार्पण राष्ट्रीय सलाहकार परिषद् की अध्यक्षा और भारतीय राष्ट्रीय कांग्रेस की अध्यक्षा श्रीमती सोनिया गाँधी ने किया। इस अवसर पर हमारे प्रधानमंत्री श्री मनमोहन सिंह भी उपस्थित थे। संयुक्त प्रगतिशील गठबंधन के साझा न्यूनतम कार्यक्रम का एक मुद्दा सूचना प्रौद्योगिकी का वहनीय कीमत पर प्रसार करना भी था। इस सीडी को जारी करते समय तत्कालीन संचार एवं सूचना प्रौद्योगिकी मंत्री श्री दयानिधि मारन ने कहा था ''मुझे पूरा विश्वास है कि कंप्यूटर के लाभ आम जनता तक पहुँचाने हों तो इसका उपयोग उनकी अपनी मातृभाषा में किया जाना अनिवार्य है। इसके लिए आवश्यक फोन्ट और भाषा के सॉफ़्टवेयर उपकरण लोगों को आसानी से उपलब्ध होने चाहिए। मुझ उम्मीद है कि हिंदी भाषा से संबंधित ये सॉफ़्टवेयर उपकरण तथा फोन्ट नि:शुल्क उपलब्ध होने से आम जनता को भी कंप्यूटर का प्रयोग करने के नए अवसर मिलेंगे।''

उन्होनें यह भी कहा कि सूचना प्रौद्योगिकी का आम जनता तक तब तक प्रसार नहीं किया जा सकता जब तक कि भाषाई फोन्ट और सॉफ़्टवेयर उपकरण मुफ्त और मुक्त उपलब्ध न हों। जो प्रौद्योगिकी वहनीय होगी वही भारतीय लोगों की रचनात्मक क्षमता को उन्मुक्त करने की संभावना रखेगी जैसा कि विकसित देशों का अनुभव रहा है।

इस सीडी में दिए गए फोन्ट और सॉफ़्टवेयर उपकरण भारत सरकार के सेंटर फॉर डेवलपमेंट ऑफ एडवांस्ड कंप्यूटिंग (सी डेक) ने विकसित किए हैं, जिसमें निजी क्षेत्र के मोड्यूलर इन्फोटैक, साइबरस्पेस मल्टीमीडिया, प्रिया इन्फोर्मेटिक्स, आईआईटी हैदराबाद, कैडग्राफ डिज़िटल सिस्टम, सी के टैक्नालॉजी, प्रोलोजिक्स और सॉफ्टव्यू आदि ने सहयोग दिया है। यह सीडी सी डेक पूना एवं सी डेक नोएडा द्वारा वितरित की जा रही है। यदि आपको यह सीडी नहीं मिलतीं तो घबराने की कोई बात नहीं है। ये सभी उपकरण http://www.ildc.gov.in या http://www.ildc.in वैबसाइट से से मुफ्त डाऊनलोड किए जा सकते हैं यहाँ से आप इस विषय की अपनी समस्याओं का निराकरण भी कर सकते है।

इस सीडी में क्या है?

यह सीडी खुलते ही सबसे पहले आपके ऑपरेटिंग सिस्टम को पहचानती है, फिर उसके अनुसार उत्पादों के मेनू दिखाती है। उदाहरण के लिए, यदि आपका ऑपरेटिंग सिस्टम विंडोज़ 98 है तो यह सीडी ओटी फोन्ट और ड्राइवर नहीं दिखाएगी। यदि आपका ऑपरेटिंग सिस्टम विंड़ोज एक्सपी या विंडोज़ 2000 है तो यह सीडी टीटी फोंटों के साथ ओटी फोंट और उनके ड्राइवर भी दिखाएगी, और यदि आपका ऑपरेटिंग सिस्टम लिनक्स है तो यह सीडी केवल उन्हीं सॉफ्टवेयरों को दिखाएगी जो लिनक्स पर चल सकते होंगे

ओटी फोन्ट या ड्राइवर लोड करने से पहले विंडोज़ में हिंदी को सक्रिय करना जरूरी है। जिसकी विधि सीडी में सबसे पहले ही दी गई है। आपके सुलभ संदर्भ के लिए हम यहाँ भी हिंदी को सक्रिय करने की विधि दे रहे हैं

विंडोज़ 2000 में हिंदी (भारतीय भाषाओं) को सक्रिय करने की विधि

1. 'Control Panel' में जा कर 'Regional Options' को क्लिक करें।
2. General Tab में जा कर Indic को चुनें, फिर Apply को क्लिक करें।
3. फिर Input Locale टैब को दबा कर Add को क्लिक करें।

4. अपने पसंद की भाषा (यानि हिंदी या मराठी, गुजराती, पंजाबी, तमिल कोई भी) और कुंजीपटल (Hindi Traditional या Inscript-Devnagari) को चुनें और OK करें।
5. 'Regional Options' के OK पर क्लिक करें, कंप्यूटर को दुबारा चलाएँ।
6. कंप्यूटर को दुबारा चलाएँ। स्टेटस बार में एक छोटा सा EN लिखा दिखेगा। इसे क्लिक कर HI चुनें और हिंदी में टाइप करना शुरू कर दें। जब अंग्रेजी में टाइप हो तो पुन: EN चुनें और अंग्रेजी में टाइप करना शुरू कर दें।

विंडोज़ एक्सपी में हिंदी (भारतीय भाषाओं) को सक्रिय करने की विधि

1. 'Control Panel' में जा कर 'Date, Time' Language and Regional Options' पर क्लिक करें।
2. इसमें Regional and Language Options चुनें।
3. 'Language' टैब को क्लिक करें तथा 'Install Files for Complex Script' नामक बॉक्स को चुनें और Apply बटन को दबा दें।
4. यह विंडोज़ एक्सीपी की सीडी माँगेगा, विंड़ोज एक्सपी की सीडी को सीडी ड्राइव में डालें तथा संस्थापन होने दें।
5. अब 'Details' टैब पर क्लिक करें तथा Settings टैब में Add को क्लिक करें।
6. अपने पसंद की भाषा और की बोर्ड को चुन कर OK करें।

कंप्यूटर को दुबारा चलाएँ, स्टेटस बार में एक छोटा सा EN लिखा दिखेगा, इसे क्लिक करने पर HI चुनें और हिंदी में टाइप करना शुरू कर दें। इसका डिफाल्ट कुंजीपटल इंस्क्रिप्ट है।

इस सीडी में आम जनता के लिए बहुत कुछ है, चाहे आप व्यापारी हों, सरकारी कर्मचारी, किसी निजी क्षेत्र में काम करने वाले हों या कोई गृहणी, छात्र हो या किसान, यह सीडी अंग्रेजी में उपलब्ध तकनीकों के हिंदी रूप उपलब्ध कराती है और वह भी बिल्कुल मुफ्त क्या है इस सीडी में, यदि आप विंडोज़ एक्सपी पर हैं तो इसमें सब कुछ है।

1. अगर आप हिंदी में टाइपराइटर पर टाइप करना जानते हैं तो इसमें जिस्टओटी टाइपिंग टूल है। जिसकी मदद से मंगल फोंट (यूनिकोड आधारित ओपन टाइप फोंट) में अपने परिचित टाइपराइटर कीबोर्ड से भी हिंदी में टाइप कर पाएंगे, जिसे आप हिंदी भाषा के यूनिकोड आधारित कीबोर्ड ड्राइवर के नीचे दिए इंस्टाल बटन को क्लिक करके इंस्टाल कर सकते है।
2. यदि आप वरिष्ट अधिकारी हैं या आपका काम टाइप करना नहीं, बल्कि कभी कभार थोड़ा बहुत संपादन करना भर है, तो जिस्टओटी टाइपिंग टूल में फोनेटिक कुंजी पटल की सुविधा मौजूद है जिसमें अंग्रेजी अक्षरों के आधार पर हिंदी में टाइप कर सकते हैं।
3. यदि आप हिंदी में टाइप करना सीखना चाहते हैं तो इसके लिए इस सीडी में आसान नाम हिंदी अंग्रेजी टाइपिंग शिक्षक दिया गया है, जिसके इस्तेमाल से आप घर बैठे हिंदी टाइपिंग सीख सकते हैं और वह भी मात्र एक सप्ताह में।
4. यदि आप 'मंगल' फोंट से खुश नहीं हैं और ऑफिस एक्सपी में दिया एरियल यूनिकोड एमएस फोंट भी आपको पसंद नहीं है तो कोई बात नहीं, इस सीडी में 160 आपेनटाइप फोंट दिए गए हैं, उनका इस्तेमाल करें, इन्हें आप जिस्टओटी फोंट इंस्टालर, मोडयूल और कैडग्राफ के ओटी इंस्टालर तथा साइबरस्केप मल्टीमीडिया के फोंट इंस्टालर से इंस्टाल कर सकते हैं ये सभी फोंट हिंदी भाषा के यूनिकोड आधारित ओपन टाइप फोंट के नीचे दिए इंस्टाल बटन को क्लिक करके इंस्टाल किए जा सकते है।
5. यदि आपको अनुवाद करने की जरूरत पड़ती है या आप अंग्रेजी शब्दों के हिंदी अर्थ जानना चाहते हैं तो इसमें शब्दिका नामक भारत सरकार की प्रशासनिक और बैंकिग शब्दावली दी गई है, साथ ही डिक्शनरी के नाम से सामान्य शब्दकोश भी दिया गया है, इन्हें अंग्रेजी हिंदी शब्दकोश के नीचे दिए गए इंस्टाल बटन पर क्लिक करके इंस्टाल किया जा सकता है।

6. यदि आप महँगे एमएसऑफिस या ऑफिस एक्सपी को नहीं खरीदना चाहते तो इसमें हिंदी में भारतीय ओपनऑफिस है जो एमएसऑफिस जैसा ही बढ़िया है, इसमें आप हिंदी में ही नहीं दूसरी भाषाओं में भी काम कर सकते हैं।
   i. इसमें एमएस वर्ड का विकल्प 'टेक्स्ट लेखपत्र' है जिसमें टाइप करने के साथ साथ मेलमर्ज आदि उन्नत कार्य भी कर सकते है। सबसे महत्वपूर्ण यह कि आप पीडीएफ फोर्मेट भी बना सकते है।
   ii. इसमें एमएसएक्सेल का विकल्प 'स्प्रैडशीट' है जो सभी उन्नत प्रयोगों को उपलब्ध कराता है।
   iii. इसमें एमएस पावरप्वाइंट का विकल्प 'प्रस्तुति' है। इसके अलावा एक ड्राइंग टूल 'रेखाचित्र' भी है।
   iv. सबसे अच्छी बात यह कि इसमें एमएसऑफिस की फाइलें बड़ी सहजता से खुल जाती है।
7. यदि आप इंटरनेट का भ्रमण करने के आदी हैं तो इसमें फायरफॉक्स नामक ब्राउज़र है
8. यदि आप ई मेल का इस्तेमाल करते हैं तो इसमें कोलंबा नामक ई मेल क्लाइंट है।
9. यदि आप गपशप यानि चैटिंग के शौकीन हैं तो इसमें 'गेम' नामक मल्टीप्रोटोकॉल चैटिंग और मैसेजिंग टूल मौजूद है जो याहू और एमएसएम जैसे कई मैसेंजरों को सपोर्ट करता है, और एक ही मैसेंजर दूसरे सभी मैसेंजरों से कनेक्टिविटी उपलब्ध कराता है।
10. यदि आप इंटरनेट से संगीत या विडियों क्लिप ढूँढना चाहते हैं तो इसमें हिंदी लाइमवायर नामक शक्तिशाली टूल मौजूद है जो बड़ी तेजी से इंटरनेट से आपकी पसंदीदा फाइलें ढूंढ लाता है।
11. यदि आप किताबें पढ़ना तो चाहते हैं लेकिन ये चाहते हैं कि कोई दूसरा आपको पढ़कर सुना दे तो इस सीडी में वाचक नामक टैक्स्ट टू स्पीच सॉफ्टवेयर उपलब्ध है जो एमएस वर्ड की फाइलों को बड़ी सहजता से पढ़ कर सुना देगा, पूरी शुद्धता के साथ।
12. यदि आपके पास हिंदी में छपे लेख आदि हैं जिन्हें आप संपादित करना चाहते हैं, लेकिन पूरे लेख को टाइप नहीं करना चाहते तो कोई बात नहीं, इस सीडी में चित्रांकन नामक ओसीआर मौजूद है जो स्कैन की गई सामग्री को टैक्स्ट में बदल देगा जिसे अप बड़ी सहजता से संपादित या इस्तेमाल कर पाएँगे

यदि आप प्रकाशक हैं और हिंदी में पुस्तकें छापते हैं तो यूनिकोड और ऊपर बताए टूल आपके काम के नहीं हैं क्योंकि पेजमेकर, कार्क एक्सप्रैस आदि सार्फ्टवेयर अभी हिंदी यूनिकोड नहीं देते। तो आप क्या करे? इस सीडी में आपके लिए भी बहुत कुछ है।

क. अगर पर हिंदी में टाइप करना जानते हैं तो इसमें जिस्टओटी टाइपिंग टूल है, जो आपको रेमिंग्टन कुंजीपटल पर हिंदी में टाइप करने की सुविधा प्रदान करता है।

ख. यदि आप संपादक या प्रूफशोधक हैं और आपका काम टाइप करना नहीं, बल्कि कभी कभार थोड़ा बहुत संपादन करना या प्रूफशोधन करना है, तो इसमें फोनेटिक कुंजी पटल की सुविधा मौजूद है जिसमें अंग्रेजी अक्षरों के आधार पर हिंदी में टाइप कर सकते है।

ग. यदि आप अंग्रेजी के स्पैलचेक से ईर्ष्या करते हैं कि अभी तक हिंदी में ऐसा कोई स्पैलचेक क्यों नहीं बना तो आपकी खोज पूरी हुई। इस सीडी में हिंदी का वर्तनी जाँचक दिया गया है। इसमें आप अपने शब्द भी जोड़ सकते है।

घ. यदि आपको हिंदी में टाइप सीखना चाहते हैं तो आसान नामक हिंदी अंग्रेजी टाइपिंग शिक्षक है ही जो इंस्क्रिप्ट कुंजीपटल में टाइप करना सिखाता है।

ड. प्रकाशकों और मुद्रकों की सबसे बड़ी मांग फोटों की होती है, इस सीडी में 365 टीटी फोंट दिए गए हैं जो आपके कार्य को आकर्षक और पेशेवर रूप देने में आपके सहायक होगे, ये सभी फोंट हिंदी भाषा के टू टाइप फोंट और कीबोर्ड ड्राइवर के नीचे दिए इंस्टाल बटन को क्लिक करके इंस्टाल किए जा सकते है।

च. इसमें आपके लिए सबसे महत्वपूर्ण उपकरण है परिवर्तन, जो एक फोंट को आपकी जरूरत के दूसरे फोंट में कन्वर्ट करता है। इसमें 72 फोंटकोडों को

आपस में कन्वर्ट करने की सुविधा हैं प्रत्येक फोंटकोड लगभग 20 25 फोंटो का समर्थन करता है यानी हम लगभग 1800 फोंटों को एक दूसरे में कन्वर्ट कर सकते है।

आज आपेनसोर्स सॉफ्टवेयरों की मुहिम छिड़ी हुई है, न केवल निजी उपयोक्ता बल्कि बड़ें बड़े कारपोरेट घराने भी आपेनसोम्र सॉफ्टवेयरों की ओर झुक रहे है, ओपन सोर्स सॉफ्टवेयर यानी पूरी तरह मुफ्त और मुक्त सॉफ्टवेयर। यदि आप इन सॉफ्टवेयरों में हिंदी का प्रयोग करना चाहते हैं तो इस सीडी में आपके लिए लिनक्स समर्थक फोंट यदि आप पूरी तरह ओपन सोर्स सॉफ्टवेयरों के साथ जाना चाहते हैं तो इसमें लिनक्स समर्थक अतिरिक्त फोंटों के साथ साथ जिस्ट ओटी टाइपिंग टूल, भारतीय ओपनऑफिस, हिंदी फायरफॉक्स ब्राऊजर, कोलंबा ई मेल क्लाइंट, हिंदी लाइमवायर खोजक और गेम मल्टीप्रोटोकॉल चैटिंग और मैसेजिंग टूल मौजूद है।

ऐसा नहीं है कि इससे हिंदी कंप्यूटरी की सभी समस्याएँ सुलझ गई हैं। परंतु कंप्यूटर का 90 प्रतिशत काम जिन सॉफ्टवेयरों या उपकरणों के जरिये होता है, वे इस सीडी में मुफ्त में उपलब्ध है। इसलिए यह करना अतिशयोक्ति न होगा कि यह सीडी भारतीय सूचना प्रौद्योगिकी के स्वप्न को पूरा करने की दिशा में एक महत्वपूर्ण कदम है, यह महान अपेक्षाओं को जगाता है। यह निश्चय ही विश्व ज्ञान हासिल करने और स्थानीय जरूरतों के मुताबित ज्ञान को अनुकूलित करने और रचने को प्रेरित करेगा हिंदी का प्रयोग नाटकीय ढंग से व्यक्तिगत कंप्यूटरों की खपत को बढ़ाएगा, हिंदी समाज को प्रौद्योगिकी मित्र बनने में सहायता प्रदान करेगा।

लेखक-
श्री वेद प्रकाश
हिन्दी अधिकारी
ओरिएण्टल इंश्योरेंस
दूरभाष 011 22449612

5. Report on Interaction with Localization Research Centre (LRC) at University of Limerick, Ireland

Date : 12th November to 16th November 2007
Venue : University of Limerick, Ireland

Summary of discussions:

- Localization Research Centre of University of Limerick, Ireland has spearheaded the localization activities by imparting training, tools showcase, consultancy, standards development & involvement of industry.
- Because of similar challenges faced in language technologies within India it is strongly felt that a similar setup needs to be established in India
- It is proposed that the activities to be divided into two phases. In the first phase the activities need to be undertaken in project mode with involvement from LRC, University of Limerick & in the second phase, a separate centre to be formed to undertake larger activities inclusive of localization, linguistic resources, W3C, CLDR and others

Discussions with Prof. Reinhard Schaler

The Localization Research Centre (LRC) is the information, educational, and research centre for the localization community. Established in 1995 at University College Dublin under the Irish Government and European Union funded Technologies Centres programme

In early 1999, the Executive Board of the University of Limerick (UL) approved the establishment of the Localization Research Centre (LRC). This centre is the result of a merger between the Centre for Language Engineering at the University of Limerick (UL) and the Localization Resources Centre, formally based at University College Dublin (UCD). The LRC provides a wide range of support activities for the localization industry and the professionals working in it These activities include

- Research - Postgraduate research support
- Localization Technology Laboratory and Showcase (LOTS) - Showcase of Tools, distribution, Infrastructure, network and support for localization community, tools reviews.
- Training
- Publication - Localization Focus, - the International Journal of Localization (since 1996), Localization Directory (Yellow Pages), The LRC Reader (a collection of interesting articles from past issues of Localization Focus magazine)
- Conferences - Annual Localization Conference (since 1996),
- Awards Academic Localization Awards,

Some of the highlights

- Started 1996 "localization focus" newsletter, 1997 started graduate diploma in localization.
- Localization Research Centre gave services of localization free of charge for initial 3 years.
- Localization Research Centre participates in "Organization for the Advancement of Structured Information Standards (OASIS)" and to certain extent in "Localisation Industry Standards Association (LISA)" standards.
- University of Limerick has industry advisory board for Localization Research Centre activities

Revenue model of LRC

- National Software Directorate, Irish Government seed money for LRC for three years.
- EU collaborative projects with industry participation (done 6-8 projects)
- Funding from Irish Government for Localization related Projects.
- Giving Consultancy to the MNC's
- Through Post Graduate programmes and PhD programs & other Training Programmes.

EU funded project being executed / executed by Localization Research Centre, University of Limerick

- **IGNITE** - Linguistic Infrastructure for Localization: Language Data, Tools and Standards - A project that will establish a central and accessible repository of European linguistic resources, strongly connected to and supported by the localization industry.
- **ELECT** - European Localization Exchange Centre - Providing reliable information on best practice, facilitating easy access to know-how and technology, making available guidelines on linguistic and cultural customization, and enhancing the visibility and recognition of this industry in Europe and worldwide
- **Certified Localization Professional** - Establishing an accreditation system for the Software Localization Industry.
- **WEB-IT/EFCOT** - A web-based terminology database for Information Technology/European Forum for Computer Terminology
- **Transrouter** - The decision support tool for translation managers.
- **EUROMAP** - Promoting greater awareness of language technology in the emerging Information and Communication Society.
- **DIET** - Diagnostic and Evaluation Tools for Natural Language Applications
- Other projects proposed to the Commission of the European Communities under the E-Content, the 5th Framework and the Adapt Programmes.
- **Next generation localization centres** new project- EU funded project (approx. 16 million Euro's (Approx. 16 million Euros, 5 years duration,

25% contribution from industry partners in form of manpower, tools and technologies, linguistic resources & joint IPR sharing) with participation from

- Dublin Centre University text analysis, automated text tagging, machine translation (example statistical based)
- University College Dublin speech technologies
- Trinity College systems framework, content management system
- University of Limerick - localization
- IBM

New tie-ups by Localization Research Centre, University of Limerick

Localization Research Centre (LRC) - Malaysia

The leading Malaysian University, the Universiti Malaysia Sarawak at Kuching (UNIMAS), and the University of Limerick's Localization Research Centre (LRC) establish closer links between the two institutions to promote internationalization and localizationrelated research and teaching.

Localization Research Centre (LRC) South Africa

South Africa's largest university and one of the largest distance universities in the world, the University of South Africa (Unisa), and the University of Limerick's Localization Research Centre (LRC) announce their intention to establish closer links between the two institutions to promote internationalization and localizationrelated research and teaching

The two organizations are planning joint research projects, staff exchanges and joint funding applications. In addition, the agreement covers the establishment of a mirror site of the LRC's Localization Technology Laboratory and Showcase (LOTS), the holding of LRC South Africa events and training courses, and the initiation of joint research projects in the area of internationalization and localization

Google look to UL in search for I18N Project Manager

"The University of Limerick has been highly recommended as the premier academic institution that teaches the best minds in the localization / internationalization business," according to Cyndy Cartwright of Korn/Ferry International, who is working closely with Google to find a project manager to internationalize all of Google's products globally

Localization Research Centre establishes link with leading Brazilian University Centre GeNESS (UFSC)

The agreement covers, among other points, the establishment of a mirror site of the LRC's Localization Technology Laboratory and Showcase (LOTS), the holding of major LRC Brasil events and training courses, and the initiation of joint research projects in the area of internationalization and localization

"This memorandum of understanding lays the foundation for the cooperation between Europe's and Latin America's leading research centres. Combining GeNESS's expertise in Latin America's emerging export markets with the LRC's long-standing internationalization and localization research agenda will generate terrific opportunities for both organisations", said Reinhard Schaler, Director of the LRC at the University of Limerick

Research Programme at Localization Research Centre

The LRC carries out research and development in a number of areas of software localisation, including supporting technologies, such as language engineering University of Limerick awards PhD in localization Research. Following are some of the Research projects undertaken by students for their Masters PhD degree.

1. **The development of an Open Source Localization Package for the Localization of Open Source Software**
   Researcher: Kevin Bargary
   Programme of Study: Masters Degree in Computer Science from the University of Limerick
   Status: On-going
2. **An Investigation into Aspects of Cultural Theory and their Relevance in the Development of Localized Web Applications**
   Researcher: Patrice Fanning
   Programme of Study. Masters Degree in Computer Science from the University of Limerick
   Status: On-going, near completion
3. **e-Learning: An Alternative Approach for Universities Providing Training in Localization Tools**
   Researcher: Rafael Guzman
   Programme of Study: Masters Degree in Internet Systems from Dublin City University
   Status: Completed
4. **Traditional Chinese Internationalization Issues and Their Resolution Using ASP.NET**
   Researcher: Joanne Cheung
   Programme of Study: Masters Degree in Software Localization from the University of Limerick
   Status: Completed
5. **The development of an XLIFF Source Converter**
   Researcher: Kevin Bargary
   Programme of Study. Bachelor of Science Degree in Computer Science from the University of Limerick
   Status: Completed

6. **An Approach to Localizing an Existing Website Area**

Researcher: Patrice Fanning
Programme of Study: Bachelor of Arts Degree in Languages with Computing from the University of Limerick
Status: Complete

Training initiatives at Localization Research Centre

A. Certified Localization Professional - TILP-supported certification programme
B. Graduate Diploma in Localization Technology
C. MSc in Global Computing and Localization
D. Professional Development Courses
E. LRC Internationalization and Localization Summer School
F. LRC Reader

I) **Institute of localization professionals (TILP)** - Certified Localization Professional Has the primary aim to develop professional practices in localization globally. TILP is a non-profit organization, owned by its members and lead by a council elected at its annual general meeting

TILP represents localization industry professionals and professionals active in localization related areas These include

1 Software publishers and publishers of other material using electronic media
2 e-content providers
3 localization service providers
4 tools developers
5 trainers and educators ,including third level colleges)
6. Researchers

TILP s founder sponsors include Oracle, Symantec, Alchemy, Microsoft, Novell, L10NBridge, and VeriTest and so on.

The institute of localization professionals runs Certified Localization professional (CLP) certification courses

The Certified Localization Professional (CLP) project was originally funded by the European Union under its ADAPT initiative (1998-2000) and developed a framework for the certification of localization professionals under the leadership of the Localization Research Centre (LRC).

- CLP is a globally operating professional certification system, based on requirements established in consultation with key industry players.
- CLP offers entry-level qualifications and a career path for professionals
- CLP certification allows professionals greater mobility and provides companies with a globally operating professional standard when hiring
- TILP offers CLP accreditation to audited course and training providers

TILP - CLP broad course structure is as follows

1. The course has contents from three different streams such as linguistic, engineering and management. All these three streams are much relevant to the process of localization
2. Various levels of courses are offered level-1, level-2, level-3 and all these level courses has contents from all the three streams as mentioned above
3. Level-1 is the introductory course and part of it is offered online and part of it is offered onsite
4. The onsite course is delivered through local partner with local contents and local infrastructure
5. The course duration for level-1 can be roughly 2 months

Different streams and certification:

Professionals three streams (vertical and horizontal development)

- Engineering
  - Software Engineer
  - Quality Engineer
- Project Management
- Linguistic

Two levels of certification

- Certification by accumulation of credits
  i. Core modules
  ii. Elective modules

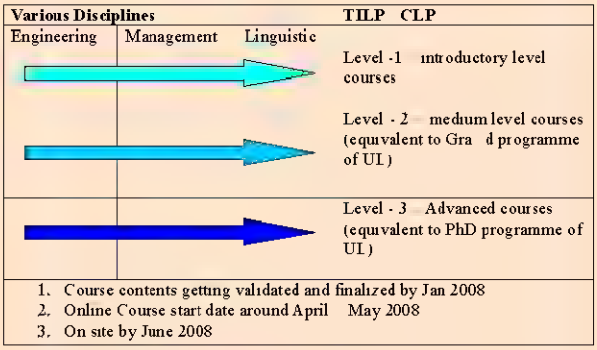
Certification:

- TILP certifies individuals as having successfully Completed a localization course provided by a TILP accredited organization.
- CLP is a part requirement for the admission as Professional member
- Course providers will pay fees for audits and certificates issued to participants. Fees will include one year associate membership

Certified localization professional CLP

- Level 1 (introductory)
- Level 2 (Diploma in localization)
- Level 3 (masters programme practicing experienced people)

Criteria for CLP localization is multidisciplinary subject consisting of Linguistic, engineering, management



Courses & Short Term Training Programmes offered by University of Limerick

The University of Limerick (UL) is offering two new programmes in Localization, designed with the support and help of the most eminent industrial and academic experts in the field. The Graduate Diploma in Localization Technology is aimed at those who want to learn how the world's digital publishers localize their products, while the MSc in Global Computing and Localization deals with the underlying scientific and business issues in localization. Each programme is now also available part-time (one-day per week over two academic years).

1) Graduate Diploma in Localization Technology

Graduate Diploma in Localization Technology is a one year taught programme. It is aimed specifically at students who might not have a technical background but who would like to become involved in localization and, above all, the technical aspects of localization. This programme is aimed specifically at students who might not have a technical background but who would like to become involved in localization and, above all, the technical aspects of localization.

Objectives

- Has a strong focus on Localization Tools and Technologies;
- Looks more closely at Internationalization, moving localization up the value chain;
- Build on the University's world-wide reputation as the premier academic institution that teaches the best minds in the localization internationalization business;
- Provide individuals interested in acquiring postgraduate academic qualifications in the areas of localization and internationalization technology with an easily accessible programme to do so.
- Provide an entry point for students aiming at MSc and PhD level qualifications in the areas of localization and internationalization

2) MSc in Global Computing and Localization

MSc in Global Computing and Localization is a one year full-time taught programme. It is aimed specifically at students who already have a background in either localization, computing, language technology, translation or related disciplines. It guides students in their research of the underlying issues in localization, with an emphasis on technical and business aspects. Students are encouraged to spend some time studying at one of UL's partner universities and work closely with industrial mentors on relevant research projects

Objectives

This programme was prepared in consultation with a large panel of academic and industry experts, and takes into account the changes in localization and internationalization practice. It caters for new and emerging requirements with a focus on high-quality research output. In addition, while it is centred around the technical aspects of localization, it recognizes the need for localization professionals to have a good understanding of international business organization.

In particular, the programme now puts a much stronger emphasis then before on:

- Internationalization requirements and their implementation;
- The critical analysis and research of translation technologies,
- The automation of the localization process in what has been called the Localization Factory;
- Business organization and international business practices.
- Build on the University's world-wide reputation as the premier academic institution that teaches the best minds in the localization internationalization business;
- Provide individuals interested in acquiring postgraduate academic qualifications in the areas of localization and internationalization with an easily accessible programme to do so, in close cooperation with industrial partners;
- Guide and support students conducting independent postgraduate research in technical aspects of localization internationalization taking into account the relevant business processes;
- Prepare students for further PhD studies.

3) Professional Development Courses

These courses will be recognized as partial fulfillment of the requirements for the Certified Localization Professional (CLP), grade 1, awarded by the Institute of Localization Professionals (TILP). All courses will be run by professional tutors. Courses will be 'hands-on'. The maximum number of participants will be 20 Courses will be delivered in state-of-the-art laboratories at the Department of Computer Science and Information Systems, University of Limerick, Ireland

On successful completion of the course, participants will receive a certificate which will be recognized by The Institute of Localization Professionals as partial fulfillment of the requirements for the Certified Localization Professional (CLP), Grade 1.

Localization Engineering

This course provides an introduction to Localization Engineering using Alchemy Catalyst case studies. Successful participants will be recognized as Certified Alchemy Catalyst Users (basic and advanced levels).

Localization Project Management

This course provides an introduction to Localization Project Management using case studies

Localization Translation and Documentation Engineering

This course provides an introduction to Localization Translation and Documentation Engineering using TRADOS case studies. Successful participants will be recognized as Certified TRADOS Users (basic and advanced levels).

Localization QA and Testing

This course provides an introduction to Localization QA and Test Automation Engineering using SilkTest case studies. Successful participants will also be recognized as automated software test developers (specifically for localized application testing)

E) LRC Internationalization and Localization Summer School

The Summer School is aimed at individuals that are researching and studying localization and localization related issues. It aims to provide individuals from different disciplines with the opportunity to experience areas beyond their own expertise.

Experience in computer programming is not required, but will need a working knowledge of computers and the windows operating system.

The LRC Internationalization and Localization Summer School will offer attendees an in-depth look at, and hands-on experience of, localization - from basic introductions to advanced concepts, and will include an intensive one day course in .NET localization

Localization Technology Laboratory and Showcase (LOTS) at Localization Research Centre

The LOTS is run by LRC and enables localization professionals and trainers to identify suitable technology faster and more efficiently. Since September 2002, localization professionals can access the LOTS laboratory free of charge. It offers the tools free of cost for non-commercial training purposes and the LOTS laboratory also helps the professionals to identify suitable technology faster and more efficiently and localisation technology developers will find it easier to establish contacts with potential customers.

The central point for learning and testing

- With this facility, researchers and developers can experiment with different technologies and operating systems, as well as language and locale settings.
- It even offers sample files, donated from various companies, to aid their research and evaluation
- LOTS encourage and facilitate high-end research

The localization industry showcase

- Central repository of commercially available tools and showcase to developers for raising awareness of the tools and technologies
- For the industry - make it easy for potential customers and users to access their products, and expand the market for localization tools and technologies.
- Users have easy, hands-on access to tools and technologies for evaluation purposes

Physical Access

- The LOTS laboratory is physically accessible to students and any professional involved in the localization industry both in Ireland and overseas

Remote Access to the LOTS Server - LOTS Online

- Most of the applications in the LOTS laboratory can be accessed by logging into the LOTS server remotely (i.e. from anywhere in the world)
- This server contains most of the software applications and sample files available in the actual LOTS laboratory.

Contents Localization Tools CD provided by LRC

LOTS CD was handed over by Localization Research Centre, University of Limerick, Ireland to Ms. Swaran Lata and Mr. M. D. Kulkarni. A small agreement in regard with usage of the same only for evaluation purpose was signed between C-DAC and LRC.

The LOTS CD consisted of the following localization tools

1 Catalyst 6.0
2 CatsCradle is a web page editor
3. Lingobit Localiser -
4. Project Open
5. WebBudget

CATALYST:

Localizer Edition is the visual solution of choice for over 80% of the world's largest software development companies. It is used to reduce localization project cycle times, improve quality and speed up delivery of software applications into international software markets

Translator/Pro Edition, it comes with a five additional Experts:

- **The Leverage Expert** turns your cost of localization into an investment and reduces the engineering and testing impact during a product update or revision
- Using the CATALYST Leverage Expert you can take translations from previous product revisions and automatically update these into new revisions, easily, automatically and safely
- **Project statistic reports** are automatically generated to identify and track all changes to your new revisions and these can be used by Project Managers to track the translation savings due to the leverage process.
- The Leverage Expert is enhanced by the addition of the Update Expert which allows the inline replacement and leveraging of individual files within a TTK
- **The Pseudo Expert** simulates the effects of translation of your software applications or Internet files and so helps the development of locale neutral applications
- **The Validate Expert** is enhanced with the addition of our Runtime Validation utility, enabling the detection of common localization bugs while your software application is actually running!
- **The Visual Comparison Expert** enables the localizer to accurately and visually determine the scope of change between software revisions in a matter of seconds

CatsCradle is a web page editor

- Fast and easy to use web page editor for professional language translators. Translate whole web-sites without having to worry about page layouts and HTML code
- Safe all-in-one web site content localization.
- Grabs all the text that requires translating from a web page, puts it into a built in editor for you to translate alongside, and then automatically integrates your translated localized text back into the web page - leaving all the sensitive HTML code untouched
- You can instantly preview your work in progress in a web browser at the click of a button There's even a real-time view where you can see the web page evolving as you translate
- CatsCradle offers full Unicode support so you can easily perform Cyrillic, Greek, Thai, Chinese and Japanese web site localization.
- Designed for translators, CatsCradle is also an ideal tool for webmasters who want to edit the text content of existing web pages. Additionally, built in support for .hhc & .hhk files means it can also be used to translate .chm help files.
- Fast and easy to use.
- Extracts all the text that requires translating from a web page, including hidden text, image alt-text etc. Just type your translations along side each line of text in the table
- When done, click 'save' and your translated text is automatically put back into the web page
- Instant preview of original and translated pages in your web browser at any time.
- Built in automatic glossary panel suggests words and phrases while you are typing - with a single key press to accept a suggestion. Helps you keep key phrases and terminology consistent throughout a project.
- Project Catalogue facility provides a single list of all pages in a project with word counts and translation status to help keep track of progress on larger projects.
- Supports hhc & .hhk help index and contents files - so can be used to translate .chm help files also.
- No pre or post processing of files - once translation has finished, you have a complete translated web site
- Safe: no worrying about fonts, page layout, hyperlinks, html code, or hidden text. It's all taken care of

Lingobit Localizer:

Lingobit Localizer extracts localizable resources from your application and makes it easy to translate, check and preview translation. When translation is ready Localizer creates localized version of your application. No source code changing is required. Localization Getting Started demo Software Localization Getting started

Localization can be done in-house or delegated to another company In the latter case, translation tasks are distributed via self-extracting localization kits with a project file (no source-code) and a 'lite' edition of Localizer for translators. Localization with Translators demo Software Localization Getting started

When you release new version of your software, old translations are automatically merged with resources from the next version and you'll only need to translate new and changed content. Translation Re-use demo Software Localization Getting started

Platforms

Lingobit Localizer is a perfect software localization tool that supports localization of several development platforms allowing translator to change text, position and other localizable parameters.

- C++, Win32 and MFC localization
- Borland Delphi & CBuilder localization
- Java localization
- XML
- Text-based files (*.ini, * cpp, *.txt)

Key Localization Features

Lingobit Localizer simplifies communication and workflow throughout the entire localization process by offering users a unified interface and productivity tools, such as Automated QA, Validation Expert, Translation Memory, etc. Lingobit's project monitoring tools give a localization manager the ability to see what is going on at each stage of the localization process, which ensures accountability, gives clarity and control to efficiently manage localization across several steps.

- Automatic Validation
- Exchange Wizard
- Translation Re-use
- Pseudo Translate
- Vesion Control
- Translation Memory
- and more..

Project-open

Web-based ERP/Project Management software for organizations with 2-200 users. It integrates areas such as CRM, sales, project planning, project tracking, collaboration, timesheet, invoicing and payments

One of the largest open-source based web applications in the world with more then 1,000,000 lines of code. It is used by more then 1000 companies in 25 countries to run their businesses.

Module Overview

- **Finance** -Tracking and evaluating company results
- **Project Management** -Internet Project collaboration and management
- **Customer Management** - A "CRM-Light"
- **Supplier Management** - Freelancers etc.
- **Human Resources Management** - Staff Employees and Skills
- **Knowledge Management** - Supporting knowledge processes
- **Data-Warehouse & Business Intelligence** - Analyzing and discovering patters in your business
- **Content & Community** - Optional modules related to creating online communities and Internet transactions around your business.
- **Systems Integration & Interfaces** - Modules that connect project-open with other systems.

Finance

The aim of the finance module is to provide the company's senior management with a real-time view to all relevant financial information of the company. For this purpose project-open provides a number of specialized modules that cover all important areas of small and medium-sized project organizations:

- Invoicing
- Timesheet management
- Travel costs
- Fixed costs
- Provider costs (via Web interface)
- Cost Center Permissions
- Financial reporting
- Export interfaces to Excel, KHK Kaufmann, ContaPlus, SQL-Ledger and SAP-FI

Project Management

The PM module integrates project-related information from all]po[modules into "project rooms" or "e-rooms", allowing you to collaborate online with providers and customers (Extranet). Sophisticated access permissions allow you to protect your business critical information

- Projects, subprojects and project task
- Project templates
- Project-related chat rooms
- Project reports and -tracking
- Risk Management
- 'Earned Value" project completion tracking
- Gantt scheduling
- Project file storage
- Project discussions
- Project news
- Incident management
- Import interface to MS-Project
- Integration with cost management
- Integration with invoicing

Customer Management

The Customer Management module ("CRM-light") unifies all functionality related to the management or customer relationships

- Customer contact management
- Integrated customer interaction history
- Online web registration
- Customer tracking
- Customer classification and status engine
- Import interfaces with MS-Outlook and ACT!

Supplier Management

The Supplier Management Module unifies all functionality related to the management of suppliers in project-oriented organizations. The strategic suppliers for this type of businesses are usually freelancers and other types of human resources, so the boundary between HR and supplier modules is blurry.

- Supplier contact management
- Integrated supplier interaction history
- Freelance skill database
- Supplier web invoice tracking
- Supplier quality module

Human Resources Management Module

The Human Resources Management Module mainly deals with the management of staff employees

- Employee payroll information
- Employee recruitment process
- Employee portraits are shown together with their office

Knowledge Management Module

- Full-text search engine for Intranet, discussions and external files
- "Expert Finder" to localize a domain expert and post questions
- "Knowledge Market" evaluates knowledge supply and use
- Extensible permission management to control the access to business critical knowledge resources

Translation Module

- Translation workflow
- Translation quality
- Translation project status reports
- Integration with invoicing module

Data Warehouse & Business Intelligence

- Multidimensional view to all corporate data
- Predefined multidimensional cubes
- Implementations available for open-source (Mondrian) and commercial (MS-SQL-Server) OLAP servers
- Predefined models for extraction of CRM key performance indicators (in collaboration with Loyalty Matrix)

Content & Community Modules

Project-open is based on the OpenACS community platform, so that all OpenACS modules are also available for project-open. These modules include

- A powerful Online store
- Classified Ads (second-hand market
- Public discussion forums
- Webmail
- Jabber Chat
- Blog & Wiki

Security

- One-Time-Password Authentication
- Build for the Internet with high-security architecture
- Role-based permission management
- Sub administration of users
- All modules are customizable (open source code)
- Secure-HTML SSL encryption

Systems Integration

- Authentication integration with corporate LDAP server
- "Multi-Company" allows to manage several separated companies with different URLs on a single server
- XML-RPC Interface

Common Characteristics

- Intuitive graphical user interface
- All modules are customizable (open source code).
- Customized (modified) modules can in general be used with the next version of po **WebBudget XT:** WebBudget XT is a world class software tool that helps language professionals and localization managers to quickly assess and translate the content of a web project.

TRANSLATE

- Support for most common tagged formats, such as HTML, SGML, XML, ASP, JSP, PHP and variations
- Support for UTF-8 encoding
- Support for most common scripting languages, such as JavaScript, vbscript
- Code-free text extraction and segmentation.
- User-friendly translation interface. Low learning curve
- Easy-to-use integrated translation memory.
- Advanced fuzzy logic, including auto-assembling feature
- TMX import export support

MANAGE

- Comprehensive text analysis in any language, including double-byte languages.
- Fully customizable reporting options.
- Generate accurate text and images quotations in a snap
- Exclusive SmartCount technology to handle delimiters and special cases.

Other Useful Features:

- Download sites with integrated new Map a site tool
- Search the site for specific keywords
- Compile the e-mail addresses found.
- Batch conversion of HTML files into RTF colored files, highlighting the translatable text.
- Easily create and exchange text extraction budgeting profiles.

The evaluation of these tools will be carried out at C-DAC, GIST lab and detailed report will be generated for future reference

Discussions with Prof Cannae - Vice President University of Limerick

We had quite interesting discussions with him in regards with taking the activity forward He gave a brief of what is the current economic situation in Ireland and what measures are being taken by the Government to increase the R&D base As regards with localization efforts in Ireland he has pointed out that in recent past MNC's are building their base for localization and Ireland seems to be a viable proposition. Following are certain reasons for the same

1. In mid 80's US companies started expanding their business in the European countries and in order to capture the local market started the localization activities in other Latin based languages.
2. US people feel at home in Ireland
3. Irish Government gave focus to localization by giving tax incentives. (Corporation tax was reduced from 40% to 10% than rest of the industry). Hence for cost effective localization Ireland was the destination for most of the MNC's
4. Because of lack of industry and opportunities most the Irish people were attracted towards other nearby countries and as a Consequence, the population in Ireland saw a dip from 7.5 million to 3.5 million. However in recent past, because of Government measures the same has now increased to 4.0 4.5 million.

Irish Government wants to give more focus on Research (especially PhD level work) rather than shifting the base to the manufacturing industry. (Since the manufacturing industry requires major infrastructure and cheap manpower

Discussions with Microsoft's Martin Orsted on localization "Challenges and issues"

Martin Orsted : Senior Group Manager, International product engineering, Ireland

He gave brief introduction of the localization activities happening in Microsoft, Dublin. He touched upon various aspects of localization and informed that he is responsible for localization of Microsoft Office Suite in almost 30 different languages & his team is poised to undertake localization for 100+ languages of world This includes 8 Indian languages as well

He also touched upon various requirements of localization and tools being developed in-house by Microsoft for the same He debated the current standard of XLIFF and indicated that it does not suit the MS localization requirements. They have developed LCX, a similar standard for the same

He also talked about various skill sets required for linguistic, engineering and management discipline at the skill levels to perform various tasks of localization (such as translation, validation, functional as well as linguistic testing, management and so on).

He gave example of waste of resources because of political issues in the given country for which localization was being undertaken.

It is expected that the localized version of the Office suite needs to be immediately released within 2-3 days after release of English version in US for Spanish, German, French & Dutch languages

He informed that in order to undertake localization for new language MS expects a minimum installed user base of 3 million licenses

He also informed that MS has undertaken localization at the instance of the Government in some countries (who has paid money to them) for the languages not covered by MS. MS looks forward to Governments for endorsement of the terminology used for localization

Following aspects need to be looked upon while gearing up for the localization

A) User of more controlled English (careful usage of Verbs, No long sentences, Simple sentences)

B) Linguistic reviews
   a. Bullet proofing source language
   b. Controlled English
   c. Identify key terms
   d. Localize in context
   e. Protect tags
   f Remove true repeats
   g. Recycle previous translations
   h. Glossaries

C) Some issues with localization
   a. Hard coded strings
   b. String dependencies
   c. String length limitations
   d. Character corruption
   e. Illegal characters
   f. Concatenating strings
   g. Un-documented placeholders
   h. Static dialogs and poor DAL implementation
   i. Don't forget the regional settings
   j. Pseudo localization
      i. Adding randomness to pseudo localization
      ii. Hybrid pseudo
   k. Approx. 3000 localization files language for Microsoft Office product
   l. 30 language localization go parallel with bi-weekly updates
   m. Move to automation and Standards based on XML
   n. Microsoft standard LCX similar to XLIFF
      i. Transforming data
      ii. Noise reduction, repeats from component based localization to resource based localization
      iii. Working in parallel, localizing and bug fixing the same file and the same time
      iv. Multiple localizer working on the same file
      v. Localization outsourced file management in-house
   o. Components of localization system -
      i. Track changes
      ii. Generate files
      iii. Update files
      iv. Concentrate file management in one place
      v. EAR Eliminate, Automate, reallocate
      vi. Contextual localization
      vii Localization in cloud
      viii. Crowd sourcing collaborate approach community effort.
   p. Testing - Advanced Test Script Frameworks Motifs, MAVI (Moving towards AI)
   q. Language Interface Pack (LIP) with community glossary
   r. EAMT European Association for Machine Translation
   s. Linguistic Quality : (retains meaning irrespective of age, literacy, region and so on)
      i. Combinatorial analysis
      ii. Orthogonal array
      iii. Directed, Ad-hoc, Ad-hoc-Directed

D) Microsoft's 'XP Lite' Microsoft's initiatives in Asia.

- Microsoft's 'XP Lite' - Low-Cost Version of Operating System sold in Asia
- The Starter Edition will ship on new, low-cost desktop PCs in Thailand, Malaysia and Indonesia. Two more countries in the five-country pilot program would be announced later this year following further discussions with governments with special pricing

- The key features of the new software would be "localized" help features, country-specific wallpapers and screensavers, and "preconfigured settings" for features that might confuse novices

Localization Standards at International levels

Major localization vendors, backed by the Localization Industry Standards Association (LISA) and The Organization for the Advancement of Structured Information Standards (OASIS), have agreed on open XML based standards for storage and exchange of data in the localization process:

A) Translation Memory eXchange (TMX) file format for exchanging translation memory data

B) TermBase eXchange (TBX) format for terminology exchange

C) XML Localization Interchange File Format (XLIFF) for extracting and storing localisable data in a common file format

D) Open source applications use GNU Gettext and the Portable Object (PO) file format.

E) SRX 1.0 Spécifications - Segmentation Rules eXchange format (SRX).

a. The purpose of the SRX format is to provide a standard method to describe segmentation rules that are being exchanged among tools and or translation vendors

F) TR29

a. Text Boundaries - describes guidelines for determining default boundaries between certain significant text elements: grapheme clusters ("user-perceived characters"), words, and sentences. For line break boundaries, see Unicode Standard Annex #14 (UAX #14), "Line Breaking Properties."

G) GMX_V1.0 Global Information Management Metrics eXchange (GMX)

a. GMX-V adopted as an OSCAR standard for the globalization industry on February 26, 2007

b. GMX-V provides a variety of statistics related to word and character counts that can be used to precisely quantify the amount of text (of various types) in a document. While it was designed with localization tasks in mind, it may be used in any field where precise, standardized quantification of text is needed.

c. The three components of GMX are

i. Volume (V)

1. Global Information Management Metrics Volume addresses the issue of quantifying the workload for a given localization or translation task. This is often commonly referred to as word counts. Word counts, however, do not convey the true range of possible statistics that can be used to assess the cost of localizing a document. Global Information Management Metrics Volume is a more precise definition of the statistics necessary for translation billing and sizing purposes

ii. Complexity (C) (proposed)

1. GMX-C will quantify the complexity of translation tasks. This format has not yet been defined

iii. Quality (Q) (proposed)

1. GMX-Q will specify the quality requirements for translation tasks. This format has not yet been defined

H) Internationalization Tag Set - ITS 1.0

a. Defines data categories and their implementation as a set of elements and attributes called the Internationalization Tag Set (ITS). ITS is designed to be used with schemas to support the internationalization and localization of schemas and documents. An implementation is provided for three schema languages: XML DTD, XML Schema and RELAX NG

Presentation by Ms. Swaran Lata

Ms. Swaran Lata touched upon various activities undertaken by TDIL Programme in giving boost to the Indian language technology development. The presentation covered an overview of activities carried out by TDIL since its inception, new activities carried out in consortia mode, various projects which are funded by TDIL, standardization activity, participation in international standards such as Unicode, W3C, IETF, ICANN, etc

Her presentation also covered

- Challenges involved in Localization for the 22 officially recognized languages with diverse scripts and more than one language based on a script & languages having more than one scripts
- The expectations from this particular visit, especially to explore possible tie up with Localization Research Centre, University of Limerick
- The objectives of the new proposed centre "National Localization Research & Resource Centre"
- Complexity of supporting Indian Languages w.r.t. Input mechanisms, rendering issues etc
- Need to establish languages with other countries for Inter-counters common languages such as Bengali, Tamil, and Nepali etc.

- The National Roll-out Initiative to promote localization by bringing out CDs. Contains BIPK in the 22 Indian Languages.
- The interface with industry which led to availability of Operating systems, Databases, applications in Indian Languages.
- To work closely with e-governance currently under implementation in the country.

Presentation by M.D. Kulkarni

Shri M D. Kulkarni covered language technology Research & Development activities undertaken at GIST labs of C-DAC, Pune. He also touched upon the importance and involvement for standards development, linguistic diversities, complexities in Indian languages (script and languages), various tools and technologies available for common man through the launch of free software and fonts CD

He has also showcased the localization framework being developed and deployed for various applications such as Railways, Banking, Insurance, e-Governance and so on

He opined a bi-lateral co-operation with LRC of University of Limerick for setting up a similar localization Centre and LOTS laboratory in India which will boost the language computing and catalyst the process of content generation in Indian languages.

He also stressed up the point that unless otherwise there are contents in Indian languages, it will be difficult to bring out technologies such as spellcheckers, grammar checkers, thesaurus, Machine Translation system and so on

Contributors to the LOTS Satellite Distribution

- www alchemysoftware ie - Alchemy Catalyst (visual localization solution)
- www stormdance net - CatsCradle (web site localization tool)
- www.passolo.com - PASSOLO (visual localization solution)
- www project-open com - Project Open (project management and workflow)
- www sdl com- SDL & TRADOS tools suite (extended LOTS Satellite Distribution only; visual Localization solution and translation memory)
- www.webbudget.com- WebBudget (web site localization tool)

Important Websites / Resources:

1. www gilc info limerick declaration pdf - GILC Global initiative for local computing
2. www electonline org
3. LOTS online
4. www igniteweb org - The IGNITE project is an EU-funded project that began in April 2005 and will run until April 2007. IGNITE will pool together linguistic infrastructure resources and provide convenient access and a market place for them
5. Localization Focus tools review available in "localization focus" and "multilingual computing"
6. www ethnologue com
7. http www oasis-open org/home index php
8. www localisation.ie - The Localization Research Centre (LRC)
9. www tilponline org - The Institute of Localization Professionals (TILP)
10. www gala-global org - The Globalization and Localization Association
11. www lisa org - The Localization Industry Standards Association
12. www unicode org - The international character encoding consortium
13. www.oasis-open org - The XML consortium developing open standards (XLIFF,Trans-WS)
14. www eeel-online org - eContent project
15. http www project-open org

One Laptop per Child - Nicholas Negroponte

Founder and chairman of the One Laptop per Child non-profit association. He is currently on leave from MIT, where he was co-founder and director of the MIT Media Laboratory, and the Jerome B. Wiesner Professor of Media Technology

A graduate of MIT, Nicholas was a pioneer in the field of computer-aided design, and has been a member of the MIT faculty since 1966. Conceived in 1980, the Media Laboratory opened its doors in 1985. He is also author of the 1995 best seller, Being Digital, which has been translated into more than 40 languages.

In the private sector, Nicholas serves on the board of directors for Motorola, Inc. and as general partner in a venture capital firm specializing in digital technologies for information and entertainment. He has provided start-up funds for more than 40 companies, including Wired magazine.

Localization will get boost through the above scheme as there will a huge demand for localized applications, tools and even localized contents.

General discussions with University of Limerick

The localization Research Centre, University of Limerick has given the "Associate membership" for one year free of cost to both Mr. Mahesh Kulkarni and Ms. Swaran Lata

Session on "Introduction to localization" by Prof Reinhard Schaler which was more focused towards the L10n testing and quality assurance.

- Testing needs to be done on three different levels such as
  - Linguistic (cultural issues, translations, locales, etc)
  - Layout (cosmetic appearance) especially for various dialog, menu items and various resolution devices
  - Functional - critical requirement for web pages localization, functionality to be retained under different operating environments and browsers
  - Test management also is an important aspect of the testing
  - Testing engineers and the localization engineer's needs to work hand in hand otherwise effective localization will not be possible
  - The testing time required for localized product is multiple of languages in which it is localized However, expectation is that the localized product testing should be done in shortest span of time
  - Testing also needs to be done in real time environment
  - Testing needs to be automated, there are various tools available for the same

Four major areas of possible collaborative efforts with Localization Research Centre and the newly proposed centre in India:

- Tools development distribution: adapting localization tools for Indian languages & distribution rights at lower cost.
- Standardization: access to standards participation in localization standards development such as LISA, OASIS etc.
- Awareness: publications, workshops, conferences
- Joint PHD programme with University of Limerick
- Certified Localization Professional Training collaboratively with LRC and TILP.

Objectives of the Proposed National Localization Research and Resource Centre (NLRRC)

The NLRRC is proposed to be set up in two phases : Phase I Implementation through the project for initiating activity and also for devising a suitable model of the proposed Centre including draft MOU between NLRRC and LRC, Ireland in above areas. Also planning to put up physical infrastructure for the Centre and obtaining necessary Govt. approvals. Project phase duration may be 18 months

Phase II the actual implementation of the NLRRC. The objectives of the proposed NLRRC are :

A) Standards & Certification

Phase I

1. Consolidation & Integration of existing activities of W3C Indian office
2. Activities covering following standards :
   - W3C
   - Storage
   - Font
   - Keyboard
   - Rendering engines
   - Transliteration
   - Ground work for certification
   - Membership of various localization standardization bodies

Phase II

- Advanced W3C activities
- Deploying certification scheme for language technology solutions

B) Localization Tools & Technology Development and Demonstration

Phase I

- Identification of various tools for localization, evaluation, distribution and licensing policies

Phase II

- Setting up of LOTS

C) Technology Support to the application providers

Phase I

- Best practices, guidelines, testing methodologies to be made available on the web
- Awareness programmes, workshops, conferences on internationalization and localization
- Development of standardized templates and initiation of linguistic resource creation.

Phase II

- Linguistic resources, tools and tools training, publications

D) Training on localization at various levels

Phase I

- Adaptation of course curriculum as per Indian language localization requirements
- Initiation of Level 1 CLP Programme

Phase II

- Post Graduate Programmes and joint PHD Programmes.

E) Consultancy, Education & Outreach

Phase I

- Deploying web portal inclusive of showcase of tools and technologies and integration of ILDC.

Phase II

- Case studies, consultancy and outreach will be implemented in Phase II.

F) Contribution to common locale data repository (CLDR)

Phase I

- To evolve mechanism for participation in CLDR by collaborating with State Govts. and CIIL, Mysore

Phase II

- Development of vetting mechanisms before finalizing contributions to CLDR

G) Linguistic Resource Development & Dissemination

Phase I

- Identification of linguistic resources and developing standards, frameworks, guidelines for Linguistic Resource Development.

Phase II

- Building of Resources, validation and mechanisms for distribution

H) E-Content proliferation in Indian Languages

Phase I

- Content Development and management tools

Phase II

- Hosting of content and developing distributed models of content generation.

Courtesy

Ms Swaran Lata
Director-TDIL Programme
MC&IT, Govt. of India
slata@mit.gov.in

Mahesh D Kulkarni
Programme Coordinator
C-DAS, GIST, Pune
mdk@cdac.in

6. हिंदी कंप्यूटरी - एक समीक्षा

''आप अपना पत्र या लेख टाइप करते हैं, तो हिंदी में क्यों नहीं, जब आप ई मेल भेजते हैं, तो हिंदी में क्यों नहीं, आप कंप्यूटर पर फिल्म देखना चाहते हैं तो फाइल खोलने के लिए हिंदी का प्रयोग क्यों नहीं करते, आप मोबाइल पर एसएमएस भेजते हैं तो हिंदी में क्यों नहीं। आप अपनी हिंदी को बेहतर बनाने के लिए इलेक्ट्रॉनिक या ऑनलाइन हिंदी शब्दकोश या समांतर कोश का प्रयोग क्यों नहीं करते। अपनी अंग्रेजी को प्रभावपूर्ण बनाने के लिए इलेक्ट्रॉनिक या ऑनलाइन अंग्रेजी-हिंदी शब्दकोश क्यों नहीं देखते। जब आप कंप्यूटर खरीदते हैं तो हिंदी की बोर्ड क्यों नहीं खरीदते, आदि आदि।

यदि मैं उपरोक्त सवाल किसी अच्छे-खासे पढ़े लिखे हिंदी भाषी आदमी से पूछूँ तो वह कहेगा कि

एक हिंदी में ये सब काम करना संभव ही नहीं है,
दो यदि संभव है भी तो हिंदी में यह सब करना महँगा पड़ता है।
तीन हिंदी में करने में समय ज्यादा लगता है, और
चार दूसरे कंप्यूटरों पर मेरी मेल खोलने या उनकी मेल अपने कंप्यूटर पर खोलना काफी दिक्कत भरा है।

लेकिन क्या यह सच है? नहीं, कतई नहीं''

ये शब्द वेद प्रकाश की पुस्तक 'हिंदी कंप्यूटरी' के फ्लैश पर छपें हैं जो अपने आप में हिंदी के महत्व और स्थिति की बड़े प्रबल ढंग से घोषणा करते है। श्री वेद प्रकाश की पुस्तक 'हिंदी कंप्यूटरी' एक महत्वपूर्ण पुस्तक है। यह पुस्तक भारत में सूचना प्रौद्योगिकी में अंग्रेजी के वर्चस्व के कारण पैदा हुए डिजिटल विभाजन को तोड़ने की एक महत्वपूर्ण कोशिश करती है। भारत में सूचना प्रौद्योगिकी का आम जनता तक प्रसार तब तक संभव नहीं है जब तक की यह प्रौद्योगिकी हिंदी और दूसरी भाषाओं में उपलब्ध न हो।

आज कंप्यूटर के लगभग सभी अनुप्रयोगों के लिए हिंदी टूल उपलब्ध हैं, और भारत सरकार द्वारा हिंदी सॉफ्टवेयर टूल की सीडी लॉच करने के बाद तो वे मुफ्त में भी उपलब्ध है। लेकिन जहाँ तक हिंदी भाषी लोगों का सवाल है, आम लोगों की तो बात ही क्या, विश्वविद्यालयों के हिंदी प्राध्यापकों और सरकारी कार्यालयों के हिंदी अधिकारियों को भी इन उपकरणों की जानकारी नहीं है।

ऐसे में वेद प्रकाश की यह पुस्तक एक ऐसे गैप को भरती है जिसकी जरूरत काफी लंबे समय से महसूस की जा रही थी। यह पुस्तक उन सवालों पर विस्तार पूर्वक विचार करती है जिन्हें हम जानना तो चाहते हैं। जैसे जब हम अंग्रेजी में कोई सामग्री किसी को भेजते हैं तो फोंट की समस्या नहीं आती तो हिंदी में ई मेल करते समय या फ्लॉपी द्वारा सामग्री भेजते समय फोंट की समस्या क्यों आती है? अंग्रेजी में तो एक ही कीबोर्ड है जिसे सीख लेने पर कोई भी अंग्रेजी में टाइप कर सकता है तो हिंदी में बहुत सारे कीबोर्ड क्यों हैं और हम कौन सा कीबोर्ड सीखें? अगर मैं हिंदी में ई मेल करूँगा तो क्या अंग्रेजी वाले दूसरे कंप्यूटरों पर उसे पढ़ा जा सकेगा? बल्कि यह भी कि क्या कंप्यूटर पर हिंदी में काम किया जा सकता है? क्या केवल पत्र टाइप करने का काम ही या हम टेबल बनाने, प्रैजेंटेशन बनाने और डेटाबैंक बनाने का काम हिंदी में भी कर सकते है? अगर हाँ तो इसमें कितना खर्चा आएगा? क्या कहा एकदम मुफ्त, नहीं हो ही नहीं सकता। आप तो मज़ाक कर रहे हैं..नहीं, यह मज़ाक नहीं है। आज कंप्यूटर पर हिंदी में काम करना बहुत आसान है और उसके सभी टूल मुफ्त में भी उपलब्ध है।

पुस्तक का पहला अध्याय 'सूचना प्रौद्योगिकी और हिंदी समाज' हिंदी भाषी लोगों की आशंकाओं को दूर कर उन्हें सूचना प्रौद्योगिकी का मित्र बनने के लिए आमंत्रित करता है, वे कहते है ''आज हिंदी कंप्यूटरी सूचना प्रौद्योगिकी में विस्फोट के लिए तैयार है। इसके लिए न उपकरणों की कमी है, न साधनों की। जरूरत है तो सिर्फ इस बात की कि हिंदी समाज और इसका प्रबुद्ध वर्ग, खासकर सरकारी हिंदी विभागों के लोग और विश्वविद्यालयों के हिंदी विभागों के लोग इनके प्रति जागरूक और संवेदनशील हो।'' (पृ. 24)

भारत के डिजिटल विभाजन को रेखांकित करते हुए वे अगले अध्याय 'हिंदी समाज के लिए सूचना प्रौद्योगिकी क्यों आवश्यक है?' में कहते हैं ''यदि हम यह चाहते हैं कि हम सूचना क्रांति के केवल बाजार, उपभोक्ता और ग्राहक ही बनें, उसकी उन्नति में योगदान भी दें, उसकी समृद्धि से लाभ भी उठाए तो हमें इससे जुड़ना होगा, यह जुड़ाव कंप्यूटर या सॉफ्टवेयर इंजीनियरों के रूप में तो होगा ही, एक बैंकर, बीमाकर्ता, साहित्यकार, पत्रकार, अधिकारी, वकील, न्यायाधीश, मनोरंजनकर्ता, संगीतकार, कलाकार, किसान, मज़दूर, दूधवाला, रिक्शावाला, चायवाला, कहने का अभिप्राय है कि हर स्तर, हर पेशे के स्तर पर भी होना होगा।'' (पृ. 25) वे आगाह करते हैं कि ''अगर सूचना प्रौद्यागिकी से यह जुड़ाव केवल अंग्रेजी के जरिए होना जारी रहा, जैसा कि अभी तक हो रहा है, तो समाज का एक बहुत बड़ा हिस्सा इसके लाभों से तो वंचित हो ही जाएगा, कालांतर में इसके प्रति द्वेषभाव भी रखने लगेगा। जो अंतत: समाज के ताने बाने को छिन्न भिन्न कर देगा। समाज के विभिन्न वर्गों के बीच बढ़ते विकराल अंतर को यदि समय रहते नहीं रोका गया तो इस समाज को बिखरने से बहुत समय तक नहीं रोका जा सकेगा।'' (पृ. 27) और ''विश्व समाज आज उत्तरोत्तर सूचना टैक्नालॉजी की ओर अग्रसर होता जा रहा है अगर समाज के लोगों को अपनी मातृ भाषा में कंप्यूटर से सूचना का आदान प्रदान करना संभव हो सके तो वे इस सूचना क्रांति में अधिक सक्रिय रूप से भागीदारी कर सकते हैं। भारत के लिए यह क्रांति केवल इसलिए महत्वपूर्ण नहीं है कि हमारा समाज बहुभाषी है, बल्कि इसलिए भी कि हमारा समाज विभिन्न आर्थिक, सांस्कृतिक और सामाजिक स्तरों पर भी बँटा है। इसलिए मानव मशीन के बीच संवाद की स्थिति पैदा करने के लिए यह जरूरी है कि उपयुक्त सूचना प्रणाली और बहुभाषा प्रौद्योगिकी के उपकरणों का विकास किया जाए और वे लोागों को किफायती कीमतों पर सुलभ हों, इसके साथ ही हिंदी और अन्य भारतीय भाषाओं में कंप्यूटर उपकरणों के प्रयोग को बढ़ावा देना भी बहुत जरूरी है।'' (पृ. 28 29)

कंप्यूटर तो कोई केवल बाइनरी भाषा समझता है उसके लिए अंग्रेजी या किसी भाषा का कोई महत्व नहीं है। जरूरत इस बाइनरी भाषा के साथ हमारी भाषा का तालमेल बैठाने की होती है। और यह काम कोंडिग व्यवस्था करती है। अंग्रेजी में आस्की कोड सर्वव्यापक है तो हिंदी में कोई कोड है या नहीं। हिंदी में इस्की कोड (भारतीय मानक सूचना अंतरविनिमिय के लिए भारतीय लिपि संहिता) 1991 में भारतीय मानक ब्यूरो द्वारा मानकीकृत किया गया। यह कोड ब्राह्मी लिपि पर आधारित होने के कारण सभी भारतीय लिपियों के लिए एक समान था इस कारण एक भारतीय लिपि (जैसे देवनागरी) से दूसरी भारतीय लिपि (जैसे कन्नड़) में पूरी शुद्धता के साथ लिप्यंतरण संभव था। सभी भारतीय लिपियों के लिए एक कुंजीपटल संभव था। भाषा के वर्णानुक्रम में होने के कारण सोर्टिंग और फिल्टरिंग संभव थे। लेकिन हिंदी सॉफ्टवेयर निर्माताओं ने उसका पालन नहीं किया। नतीजतन कंप्यूटरी हिंदी में अराजकता पैदा हुई। इस कोड के बारे में विस्तृत चर्चा 'हिंदी का मानक कोड क्यों?' अध्याय में की गइ है। उद्देश्य है कि हिंदी में मानक कोड व्यवस्था लागू हो ताकि हिंदी का विकास अबाध गति से हो सके।

और यह मानकीकरण केवल हिंदी के लिए ही नहीं बल्कि विश्व की सभी भाषाओं के लिए होता है यूनिको की संकल्पना में। क्या है यूनिकोड? 'यूनिकोड : समस्याएँ अनेक समाधान एक' शीर्षक लेख में वे उत्साहपूर्वक बताते हैं 'यूनिकोड में हिंदी और दूसरी भारतीय भाषाओं को भी स्थान मिला है।.... देर आयद दुरुस्त आयद। चूँकि यूनिकोड में देवनागरी लिपि भी शामिल है इसलिए सारे यूनिकोड समर्थित सॉफ्टवेयर खुद ब खुद हिंदी समर्थक हो गए हैं, बशर्ते कि आपने उनमें यूनिकोड बेस्ड हिंदी फोंट को सक्रिय किया हुआ हो। आप किसी सॉफ्टवेयर का नाम लीजिए, आप पाएँगे कि न केवल उसमें हिंदी में काम करने की सुविधा मौजूद है बल्कि उसका इंटरफेस भी हिंदी में आ चुका है। जो अभी नहीं आए है, वे भी इस राजमार्ग पर बढ़ रहे हैं। यही नए अंतरराष्ट्रीय समुदाय के ज्ञान सूचना जानकारी के आदान प्रदान का आधार है।' (पृ. 41 42) वे आहवान करते हैं हिंदी कंप्यूटरी में मानक स्थापित करने ही होंगे। हमें हर हाल में यूनिकोड को अपना लेना चाहिए। यह हिंदी कंप्यूटरी की दुनिया में एक नए युग की शुरूआत है। अब हिंदी कंप्यूटर पर केवल पत्र टाइप

करने वाली भाषा बन कर नहीं रह सकती। इसे नई प्रौद्योगिकी के हर चरण में अपनी ताकत दिखानी है। यदि हिंदी समाज प्रण कर लें कि वह यूनिकोड व अन्य तय मानकों का ही प्रयोग करेगा तो फिर सूरज निकला ही समझो।' (पृ. 48)

जिसने भी कभी हिंदी में कंप्यूटर पर टाइप किया है या करवाया है वह निश्चित ही हिंदी फोंट की समस्या से जूझा होगा। इस समस्या के कारणों पर विस्तारपूर्वक जानना हो तो 'ये हिंदी फोंट क्या है? शीर्षक अध्याय पढ़िए। कितनी तरह के हिंदी फोंट होते हैं- पोस्ट स्क्रिप्ट फोंट, टू टाइप फोंट, अंग्रेजी फोंट, बाइलिंग्वुअल फोंट, डायनेमिक फोंट, वैब फोंट और ओपनटाइप फोंट। लेकिन सम्सया का हल क्या है? 'ओपन टाईप फोंट सब रोगों की दवा हैं, ये यूनिकोड आधारित सभी अनुप्रयोगों को समर्थन देते हैं।' (पृ.58)

एक पाठ -विंडोज़- 2000 और विंड़ोज एक्सपी में हिंदी को सक्रिय कैसे करें पर भी है। जो बहुत उपयोगी है। लेकिन सबसे महत्वपूर्ण पाठ है 'हिंदी में टाइप कैसे करें?' इस पाठ में उन्होंने रेमिंग्टन कीबोर्ड और इंस्क्रिप्ट कीबोर्ड की बात की है। इंस्क्रिप्ट की बोर्ड की वैज्ञानिकता, बहुभाषिता, तीव्रता आदि को इतने अच्छे ढंग से बताया गया है कि आप पुस्तक पढ़ते-पढ़ते ही हिंदी में टाइप सीखने को उत्सुक हो उठते है। यह अध्याय हिंदी में एक कुंजीपटल की वकालत करता है। और हिंदी सॉफ्टवेयर निर्माताओं द्वारा दुनियाभर के कीबोर्ड उपलब्ध कराने की निंदा करता है। उनका दावा है कि 'यदि हम केवल गृह कतार के अक्षरों के स्थान याद कर लें तो हिंदी के तकरीबन 70 फीसदी अक्षरों का स्थान याद हो जाता है। गृह कतार के अक्षरों को याद करने का आसान तरीका है, केवल एक पंक्ति 'ओए अइ उ परकत चट' को याद कर लेना।... इस प्रकार यदि हम चाहें तो इंस्क्रिप्ट कुंजीपटल पर मात्र 1 घंटा अभ्यास करके हिंदी में टाइप करना सीख सकते हैं।' (पृ. 101)

हमारे स्वतंत्रता सैनानियों ने एक सपना लिया था कि यूरोप की तरह भारत की सभी भाषाओं की लिपि एक हो, देवनागरी हो। इस सपने में महात्मा गाँधी, शहीद भगत सिंह, बाल गंगाधर तिलक जैसे महान नेता शामिल थे। इंस्क्रिप्ट कुंजीपटल इस सपने को पूरा करने के लिए परिवर्धित देवनागरी की कूटबद्ध करता है। यानी उन ध्वनियों को भी टाइप करने की सुविधा प्रदान करता है जो हिंदी में नहीं हैं लेकिन हमारी दूसरी भारतीय भाषाओं जैसे उर्दू, तमिल, तेलुगू आदि भाषाओं में हैं। वे ठीक ही इंस्क्रिप्ट कुंजीपटल को राष्ट्रीय एकता की ओर बढ़ता कदम कहते है 'इसलिए देवनागरी इंस्क्रिप्ट कुंजी पटल को अपनाना भारत की राष्ट्रीय एकता को लिपि एकता और भाषा एकता के जरिए सुदृढ़ करना है।' (पृ. 109)

पुस्तक का अंतिम पाठ भारत सरकार द्वारा जारी हिंदी सॉफ्टवेयर उपकरण' नामक सीडी पर है। वे कहते हैं 'भारतीय भाषाओं की कंप्यूटरी के इतिहास में वर्ष 2005 ऐतिहासिक है। इस वर्ष भारत सरकार ने ऐसा कदम उठाया है कि जिससे भारतीय भाषाओं की कंप्यूटरी के रास्ते में आने वाली सारी बाधाओं और समस्याओं का हल भाषा सीडी में उपलब्ध करा दिया। और वह भी बिल्कुल मुफ्त, जिसे आप निस्संकोच अपने मित्रों को वितरित भी कर सकते है।' (पृ. 110) इस अध्याय में वे इस सीडी में दिए एक एक उपकरण पर चर्चा करते हैं। वे गदगद होकर कहते हैं 'यह कहना अतिशयोक्ति न होगा कि यह लोकार्पण भारतीय सूचना प्रौद्योगिकी के स्वप्न को पूरा करने की दिशा में एक महत्वपूर्ण कदम है, यह महान अपेक्षाओं को जगाता है। यह निश्चय ही विश्व ज्ञान हासिल करने और स्थानीय जरूरतों के मुताबित ज्ञान को अनुकूलित करने और रचने को प्रेरित करेगा। हिंदी का प्रयोग नाटकीय ढंग से व्यक्तिगत कंप्यूटरों की खपत को बढ़ाएगा, हिदी समाज को प्रौद्योगिकी मित्र बनने में सहायता प्रदान करेगा। (पृ. 120)

अंत में मैं इस पुस्तक की भूमिका के लेखक यूपीएससी के सदस्य डॉ. पुरुषोत्तम अग्रवाल की इस बात के साथ पूरी सहमति व्यक्त करती हूँ कि 'इस किताब को पाठक हाथों हाथ लेंगे कम से कम मुझे तो इस बात का पूरा यकीन है।'

पुस्तक का नाम	:	हिंदी कंप्यूटरी
लेखक	:	वेद प्रकाश
प्रकाशक	:	लोक मित्र, 1/6588, सी-1, रोहतास नगर (पूर्व), शाहदरा, दिल्ली-110032
प्रथम संस्करण	:	2007
मूल्य	:	175/

समीक्षक
कु. स्वर्ण लता
निदेशक, टी.डी.आई.एल.
सूचना प्रौद्योगिकी विभाग

7. Appreciation for the Language CDs

"..... I appreciate the efforts of ILDC team in bringing the multi language cds. It will give the people an opportunity to learn the Indian languages and culture more easily. Pl send me a telugu language cd containg all the products offered by you....."

Shri Suresh kumar kannath.
Email : kannath-sk@yahoo.co.in

"..... I appreciate the work of CDAC, TDIL and all who involved in this great work. There is no doubt it will be very useful to all the Indians residing in India and abroad. But lots of people are not familiar in typing on their local languages. To overcome this problem, it will motivate the Indians and other enthusiastic people to use Indian languages, if CDAC and TDIL will release a typing tutor in Indian languages. My sincere wishes for CDAC, TDIL and Ministry of IT to continue its success in this IT awareness project to our country people....."

Shri Bala
Network Administrator
Gulf Pharmaceutical Industries, UAE
Email : bala@julphar.net

"..... I have been using your software by downloading it. I will request you to kindly send me the original CD, I have numbers of relative aboard and they also want the same software....."

Lt.Col Ashok Mansingh(Retd.)
Email : manjulashokmansingh@airtelbroadband.in

"..... I appreciate and thank you for the efforts to release the software/ fonts in Tamil - CD. Please send one CD, recently released, by the best efforts of our Hon'ble Minister. Thank you. I once again place on record our since thanks and appreciation for the best efforts being taken in this regard....."

Shri S.Sadasivam
E-mail : try sivasree@sancharnet.in

"..... I received the cd containing Hindi fonts, and express my thanks to you. It is indeed a great service you are rendering to Indian language dtp work....."

Shri Srinivas
Email : srivaturi@gmail.com

"..... I am glad and grateful for the compact disc of fully functional Hindi Software Tools. Now I can mail a lot of old friends in their mother tongue, and keep Hindi (atleast as little as I know) alive. I eagerly anticipate the release of Telugu Software Tools, and would be glad to help you in the process in whatever little way I can. Please do let me know if I can participate in this wonderful endeavour to bridge the so called "digital divide". Once again, my whole-hearted thanks to everyone in the team....."

Shri Muralidhar
Email : kamudhar@gmail.com

"..... Greetings!! Thanks for the initiative taken in developing Linux for the masses. I got the Hindi language tools and the utilities CD with July 2005 Issue of Linux....."

Shri Dhananjay M. Bhate

"..... I have received the CD for TAMIL, which is very useful. I thank you very much for the immediate response. My special thanks to Mr.Dayanidhi maran, the minister who is very keen to bring all language fonts. Special salute to him....."

Shri V.Sivasankar.
Email : pavai_sankar@yahoo.co.in

"..... I am impressed with the decision of distributing the free CD for Indian language for the growth of our indian language. I am also a great supporter of Hindi language here in US. I was exploring the possibility ....."

Shri Rajeev Agrawala
82 Woodbridge Terrace,
#M,Woodbridge,
New Jersey -07095.,
USA
Email : rajmadhuri@hotmail.com

"..... Sincere thanks for your help. Now Chitrankan has been properly installed and is running perfectly....."

Dr. Deepak Kaikini
Email : dakaikini@vsnl.com

"..... I got the CD lastweek on 9th of june 2005. As I go through CD It was amazing. Thanks to the efforts taken by Central Government,Officers, and Ministers. This is a very good move taken by the government to improve the Classical language. Thanks once again for the CD....."

Shri Raja Raman.R
Email : rajakkal@gmail.com

"..... aapki site bahut achchhi hai apne jo hindi ka prachar karne ka tarika apnaya hai bahut hi achchha hai isse hamare desh me bad rahi english bhasha kam hogi aur hindi badegi....."

Shri raju mishra
E-mail : krishankantr@rediffmail.com
Organisation: G Internet computer center

"..... I received the "Tamizh software tools", i am happy about it and thanks for sending it, its awesome to see tamizh fonts in my system. And its a great thing....."

Shri dhananjayan
Tamilnadu
E-mail : vigneshvaran 1984@yahoo.co.in

"..... We are in inertial systems for spacecraft mechanisms. As part of our programme we have conducted a Hindi Technical Seminar for which you have kindly supplied us with 300 copies of the Hindi Software Tool CDs on our request. As these CDs have been very beneficial to us there has been an increased demand for the same....."

Shri V.J.Sarwade,
Head, PPEG
ISRO Inertial Systems Unit
Email : vj_sarwade@vssc.gov.in

'..... I, Head Master, received the CD of Urdu software tools. We the staff are thankfull to the task, to propagate the Urdu language by your organization...."

Shri Ashfaque I. Shaikh
Head Master
M.T.P Urdu High School, Jalgaon

'..... Most respectfully and humbly i beg to state that you Department has performed a wonderful job towards the implementation of official Language (O.L) Act & Rules and progressive use of Indian Regional Languages to bring the Indian people closed together. We as the member of Indian society are very grateful to you for you eminent achievement in the positive and needful direction of time....."

Shri D.S. Singh
Gurudwara zone, Mauipur
Tel : 2256-2137

'..... Although belated, the steps taken to promote I.T. deserve applauds from all corners...."

Shri Ashok Vinayak Vaidya
Thane, Maharashtra

Hindi

राष्ट्रीय संगोष्ठी

भारतीय भाषी अभिकलन एवं प्रक्रमण के लिए शाब्दिक संसाधनों का सृजन

Punjabi

ਰਾਸ਼ਟਰੀ ਗੋਸ਼ਟੀ ਭਾਰਤੀ ਭਾਸ਼ਾਵਾਂ ਚ ਕੰਪਿਊਟਿੰਗ ਅਤੇ ਪ੍ਰਕ੍ਰਿਆਕਰਨ ਲਈ ਸ਼ਾਬਦਿਕ ਸਰੋਤਾਂ ਦਾ ਸਿਰਜਣ

LRIL-2007 Proceedings

NATIONAL SEMINAR

CREATION OF LEXICAL RESOURCES FOR INDIAN LANGUAGE COMPUTING AND PROCESSING

Sanskrit

[illegible]

Marathi

भारतीय भाषांच्या संगणन व प्रक्रियेसाठी शाब्दिक स्त्रोतांची निर्मिती

भारतीय भाषाओं में ज्ञान विज्ञान

Organizers

1. Commission for Scientific and Technical Terminology (CSTT)
New Delhi
2. Centre for Development of Advanced Computing (C-DAC)
Juhu, Mumbai

सी डैक
CDAC

Guest Editorial...

Large population of the country still remains deprived of the benefits of IT, because of the non-availability of user-friendly medium in local languages. Lexical Resources are the backbone of all the computing and processing activities. But creation of Lexical Resources has been mostly neglected by their consumers. Therefore, it is the dire need of the hour to create a platform to discuss the creation of lexical resources for language computing and processing.

Commission for Scientific and Technical Terminology (CSTT) has, as per the President's order dated 27.4.1960, the mandate to develop scientific and technical terminology for all Indian languages. **TDIL** foresees collaborative development of Indian Language lexical resources. C-DAC Mumbai under TDIL-funded jnanabhaarati project organised the National Seminar LRIL-2007 in collaboration with CSTT. We hope, using collective wisdom of experts from both the domains computer science and linguistics, for enriching the computer medium, is a right step towards improving the conditions for equitable and affordable access to computers.

Here we are publishing some of the papers presented during the seminar LRIL-2007 held at C-DAC Mumbai for wider dissemination.

Zia Saquib, *Executive Director, C-DAC Mumbai*

with assistance from **Bira Chandra Singh**, *Linguist C-DAC Mumbai*

8. Lexical Resources for Indian Language Computing and Processing (LRIL-2007) : Report

A three-day national seminar on 'Creation of Lexical Resources for Indian Language Computing and Processing' was held at C-DAC Mumbai, Juhu centre from 26th to 28th March 2007. It was jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, DIT, MC&IT, Govt. of India.

After the mellifluous recitation of a vandanaa for goddess Sarasvati, the seminar was inaugurated by lighting the traditional lamp. Prof. K. Bijay Kumar, Chairman, CSTT New Delhi, Mr. Zia Saquib, Executive Director, C-DAC Mumbai, Mr. Madhukar Sinha, Director (ICR) HRD Ministry, Prof. N. Rajasekharan Nair, Professor of Linguistics, Annamalai University, Mr. Deepak Kumar, Scientific Officer, CSTT, Mr. R. S. Rawat, Deputy Director Implementation, Home Ministery, Prof. R. K. Joshi, Visiting Design Specialist, C-DAC Mumbai, Dr. Alka Irani, Sr. Research Scientist & Head LCG, C-DAC Mumbai graced the occasion of lighting.

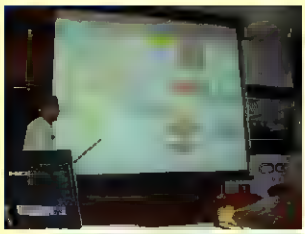
Mr. Zia Saquib pronounced a hearty welcome address to all the delegates and audience; afterwards, briefly introduced C-DAC Mumbai's leading role in R&D initiatives related to software technology, and in particular, Indian language computing. There was an invited talk by Mr. Madhukar Sinha on copyright issues with specific reference to software and lexical resources. Prof. K. Bijay Kumar delivered a very informative and lively talk on 'Principles of Evolving Technical Terminology'. Prof. N. Rajasekharan Nair, in his keynote speech, gave a comprehensive account of the lexical resource building activities in India and abroad.

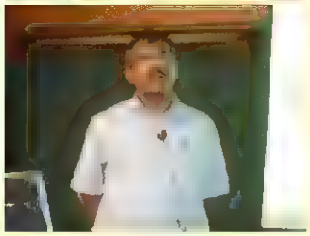
The seminar received a warm response from across the country. More than 80 scholars participated in the event. Delegates from IIT Bombay, University of Mumbai, IIIT Hyderabad, C-DAC Kolkata, University of Hyderabad, Microsoft Research India, Bangalore, Thapar University Patiala, University of Punjab Patiala, JNU New Delhi, Annamalai University Tamilnadu, Punjabi University Patiala, Pondicherry University, Banaras Hindu University Varanasi, University of Kerala, Aligarh Muslim University, Kendriya Hindi Sansthan Agra, Indlinux, Sarai New Delhi, etc. presented their research papers. Some of the delegates delivered talks on their hands-on experience also. Attendees from C-DAC Pune, Red Hat Software services Pune, etc took active part in the discussions.

In total there were ten technical sessions spread over three days. Following papers were presented in the seminar:

1. Collection, Development and Publication of Linguistic Resources Issues and Experiences with Indian Languages: Baskaran Sankaran & A. Kumaran.
2. Speech Corpora Development in Indian Languages: Dr. Shyamal Kr. Das Mandal & Arup Saha.
3. Problems of Urdu-Hindi Automatic Transliteration: Dr. Sabahuddin Ahmad & Dr. Abdul Aziz Khan
4. Statistical Analyses of Myanmar and English-Myanmar Text Corpora: Hla Hla Htay, G. Bharadwaja Kumar, and Dr. Kavi Narayana Murthy
5. Lexicography in Hindi: Challenges and Opportunities in the 21st century: Dr. Abhishek Avtans
6. Art of Hindi Dictionary Making: An Historical Exploration: Ravikant Sharma
7. The Structure of a Dialect Dictionary of Agricultural Vocabulary in Tamil: Dr. S. Raja
8. Lexicographic Traditions in India and Sanskrit: Dr. Malhar Kulkarni
9. Issues in Developing Corpus for Malayalam from Web as Source: Dr. S. A. Shanavas
10. Study of Cognates Among South Asian Languages for the Purpose of Building Lexical Resources: Anil Kumar Singh & Harshit Surana
11. Rule-based Machine Translation System using Indian Logic for Discourse Texts: Kommaluri Vijayanand
12. A Frame Work for Designing Punjabi Deconverter & Case Markers for Punjabi Language Server: Parteek Bhatia
13. On Developing Lexical Resources (with Special Reference to LRs Developed at IIIT(H)): Prof. Thakur Dass
14. Creation of Lexical Resources for Machine Translation: Dr. Rajendra Kumar Gupta
15. Causative Compound Verb Constructions: A Generative Lexicon Account: Dr. Sanjukta Ghosh and Dr. Anil Thakur
16. Handling Polysemous Particles in Multilingual Environment: Dr. Anil Thakur and Dr. Sanjukta Ghosh

17. Prosodic Elements in Composite Argumentative Connectors of Marathi: Ms. Ujjwala Joglekar
18. Evolving Translations & Terminology - the Open Source Way: G Karunakar & Ravishankar Shrivastava
19. Bricks and Mortar for Digital Resources in Indian Languages: Gora Mohanty
20. Design and development of Punjabi Spell Checker: Prof. G. S. Lehal
21. Corpus Based Statistical Approach for Stemming Telugu: M. Santhosh Kumar & Kavi Narayana Murthy
22. Morphological Analyzer for Great Andamanese Verbs: Implementing a Concatenative Template: Narayan Kumar Choudhary, Anvita Abbi, and Girish Nath Jha
23. Problems in Developing Lexical Resources for Computing: Rita Mathur
24. Multilingual Wordnet Creation with Focus on Indian Languages: Prof. Pushpak Bhattacharyya & the Wordnet Group
25. Education in Mother Tongue for the Tribals – Need of Common Lexical Resource Database and Linguistic Harmony of Indian Languages: Subodh Hansda
26. Named Entity Recognition for Telugu: P Srikanth, Kavi Narayana Murthy
27. Automatic Construction of Telugu Thesaurus from Available Lexical Resources: M.Santhosh Kumar, Kavi Narayana Murthy.



Mr. Deepak Kumar, convener of LRIL-2007 expressed his thanks and appreciation to all the delegates and urged them to be involved and cooperative in CSTT's future activities.Some exquisite posters on the significance of 'shabda' or word (designed by Prof. R. K. Joshi and his supportive team) enhanced the aesthetic beauty of the event.






The final day was graced by our special guest Mr. Som Dutt Dadheech, HOD, Human Centered Computing Division, TDIL, DIT, MCIT, Govt. of India, who affirmed positively to extend his full support for such activities in future. The seminar concluded with a panel discussion and valediction. "Transforming India into a Developed Nation with a Language of Her Own" was the topic of discussion. Dr. Dharmendra Kumar, CSTT chaired the discussion and the panel included Dr. Malhar Kulkarni, IIT Bombay, Dr. Abhishek Avtans, Kendriya Hindi Sansthan Agra, Mr. G. Karunakar, IndLinux, Mr. Anant Joshi, L&T Infotech, Mr. N R Bheda, Human Settelments Environment & Youth Centre, Prof. R.K. Joshi, C-DAC Mumbai, and Prof. Lion Luthriya, media personality.

Finally, everybody felt that it was not, in fact, the end of the event but the true beginning of further events. It was decided that further communications will be made via e-mail. Dr. Alka Irani, in conclusion, thanked and expressed her deep sense of gratitude to all for making the event successful.

Mr. Som Dutt Dadheech and participants

C-DAC Mumbai Team
(Language Computing Group)

Courtesy :
Centre for Development of Advanced Computing, Mumbai
Tel. : +91 22 26201606, +91 22 26201574
Fax : +91 22 26210139, +91 22 26232195
Web : www.cdacmumbai.in

Keynote Address

8.1 Lexicographic Tradition in India

(With Reference to Lexical Resources)

N. Raj Shekharan Nair, Annamalai University

I. INTRODUCTION

INDIA has a long and ancient lexicographic and grammatical tradition. Ya ska's Nirukta, the earliest etymological work of Sanskrit, Amarkosh of Amarasimha a traditional verse dictionary, and Nighantus belong to the dictionary tradition of India Koshas and Nighantus were mostly lists of synonyms and homonyms. In addition, the ancient grammatical works like Paaini's Aṣta.dhya:yi: in Sanskrit belonging to the 5th B.C. and Tolka ppiyar's Tolka·ppiyam of Tamil of 3rd C.B.C. provide lots of lexical information. We can say that these two and other traditional grammars were partial lexical resources. Asta dhya:yi: has accessory texts called *dhatupa:tha,* verb roots, *ganapa:tha*, lists of nominals, *linga:nu sa:sana*, gender of Sanskrit words, *phit sutras* rules on accentuation of words. In Tamil Tolka ppiyam list of classes of words are included as part of the text itself. It devotes one whole chapter for describing the meanings of lexemes or content words This chapter can be considered to contain the seeds of later lexicographic tradition of Tamil. This tradition combined with the Sanskrit Kosha tradition lead to the compilation of numerous verse Nighaṇṭus in Tamil The dictionary tradition in India owes itself to the Western Christian missionaries, which started in the 17th century. This was later followed by the native scholars. These are the beginnings of lexical resources in Indian languages

As a linguist, my thrust in my address will be to outline the different types of lexical information that should find a place in the lexical resources to be developed for Indian languages. It will be pointed out that importance must be given to the special grammatical and semantic features of Indian languages. While grammars of languages will emphasize the structural aspects like syntactic patterns, word formation, etc., which are common to classes or groups of words, lexical resources must concentrate on the formal (structural) and semantic aspects of individual words. Formal aspects will include the marking of the parts of speech (form class) of the words like noun, verb, adjective, adverb, etc., sub-categorization on the basis of inflectional classes (e.g., conjugation classes of verbs), gender marking, etc. A special feature of Indian languages is the case inflexion of nouns which is equivalent to the prepositional phrases of English Here the case selection or rection of verbs is very important for Indian languages. For example a class of involuntary physical and mental state verbs having the meanings *'be hungry', know, understand,* etc. do not take nominative case or subject in a sentence. They take only dative or fourth case, e.g., Hindi, *muje bhu:k lakti:hai*, Tamil *enakku pacikki* r̲atu, Telugu, na:ku-a:kaliga:undi 'I am hungry' (lit 'It is hungry to me') Malayalam· enikku visakkannu. Providing of case selection features of verbs for Indian languages is a must

At the semantic level meanings of polysemous and homonymous words, synonyms, antonyms, class inclusion (hyponymy), part of a whole (meronymy), etc. must be indicated In addition collocational restrictions (e.g., dog barks; horse neighs; lion roars) must also be indicated

Providing of both grammatical and semantic information must be on the basis of exhaustive study of vast corpus of each language. For this corpus development of multiple natures, is essential. Lexical resources can also study dialect variation, both regional and social (i.e., caste and religious group based). Recording of different occupational vocabularies of Indian languages is an urgent need as international terms are fast replacing the native words. This is essential to preserve the lexical wealth of Indian languages

Lexical resources will be of great use for language learning-teaching, creative writing, translation between Indian languages, creation of technical terminology, and for sociological, cultural and historical study of the different Indian language groups. The effort of C-DAC will go a long way in fulfilling the needs of different types of language users and create awareness among the common public of the multiple uses of lexical resources, apart from coordinating the works of different agencies already involved in corpora and lexical resource development

II. LANGUAGE CORPORA

India is a multilingual country. As per the census of India there are 1652 mother tongues. There are many languages, especially tribal languages, which are at the verge of extinction due to various reasons. So it is necessary to generate corpora for every Indian language. Various institutions are engaged in the preparation of corpus. The corpus developed by Shastri (1988) Shivaji University is the first corpus for Indian English called 'Kolhapur Corpus of Indian English' (KCIE).

The following is a list given by Niladri Sekhar Dash (Indian Linguistics 64: 2003) on the corpora generation projects involving Indian languages.

N Rajasekharan Nair is with CAS in Linguistics Annamalai University Annamalai Nagar 608 002 (e-mail. rajasekharan245@yahoo co in)

Agencies	Indian languages corpora
Indian Institute of Technology, Kanpur	Hindi, Nepali
Indian Institute of Technology, Mumbai	Marathi, Konkani
Indian Institute of Technology, Guwahati	Assamese, Manipuri
Indian Institute of Sciences, Bangalore	Kannada, Sanskrit
Indian Statistical Institute Kolkata	Dangla
Jawaharlal Nehru University, New Delhi	Sanskrit
University of Hyderabad, Hyderabad	Telugu
Anna University, Chennai	Tamil
MS University, Baroda	Gujarati
Utkal University, Bhubaneswar	Oriya
Thapar Institute of Engg. and Tech., Patiala	Punjabi
ER&DCI, Tridendrum	Malayalam
CDAC, Pune	Urudu, Sanskrit, Kashmiri

The above studies show the significance of such ventures.

III. SPEECH CORPORA

The scholars in many universities in foreign countries are showing keen interest to create the speech corpora for Indian languages. This seems to be a modern trend in creation of lexical resources. Recently Prof. Peter Juel Henrichsen, Centre for computational modeling of Language (CMOL), Department of Computational Linguistics, Copenhagen Business School, Denmark visited my Department (Centre of Advanced study in Linguistics, Annamalai University) and had discussions with all the faculty members. During the interaction he informed us their plan to create the speech corpora for Indian languages, especially for Tamil and a few Dravidian Tribal Languages. A colleague of mine visited the Copenhagen Business School later by invitation and interacted with the faculty members there. Now we are exploring the possibilities to have a MoU with them on the creation of the speech corpora. Niladri Sekhar Dash has highlighted the importance of speech corpora in his article on 'Speech Corpora and Text Corpora' (Indian Linguistics: 67: 2006)

Prof. Peri Bhaskara Rao, Chair, Institute for the study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign studies, Tokyo, Japan has been collecting linguistic data from the Toda Tribal people of Nilgiris, Tamilnadu for the past few years.

Dr. Christiane PILOT Raichoor, a renowned scholar in Dravidian Linguistics, LACITO-CNRS, France has been engaged in the digitization of Dravidian Tribal languages of Wayanad District, Kerala. I have been involved in the collaborative project. I have recorded the folk songs of two Dravidian Tribal communities, namely, Mullakurumar and Kurichyar. The digitization process was initiated at LACITO during November 2006 when I visited there. This will be a speech corpus, with other relevant details incorporated. This digital corpus will be a document for the scholars to do future research. The collection of speech data from other tribal communities of Wayanad district will be continued and the data will be digitized. I would like to mention here that these speech varieties are at the verge of endangerment. David Crystal in his book 'Language Death' points out that a language will die if people of that community are not using that language. The mother tongue use of many tribal communities is decreasing. Mostly in home domain they use their mother tongue. The older people who use their languages are dying due to illness, etc. The younger generation, who study in schools and colleges in Malayalam language medium, which is the major language of Kerala, use their mother tongue less frequently. In this circumstance there is all possibility that these tribal tongues will be lost soon. Hence we are recording the languages along with folk songs and digitizing them.

IV. DICTIONARIES

The lexicographic as well as computational lexicographical activities are going on in Indian Universities and Institutions in an appreciable manner. The trained linguists, lexicographers and computational lexicographers are involved in the compilation and also in the research. At the Centre of Advanced Study in Linguistics, Annamalai University I have completed along with Dr. S. Raja a UGC Major Research Project entitled 'A study of Dialect Variation in Tamil Agricultural Vocabulary'. Data were collected from all over Tamilnadu – mainly from the five major dialect regions of Tamilnadu. Both the lexical variations and phonological variations are identified. The results of this survey are presented in the form of a dialect dictionary, showing the variation. The same recorded cassettes can be used to generate speech corpora.

Dr. Christiane PILOT Raichoor, who is a Dravidologist, has been working on Badaga language spoken in Nilgiri district of Tamilnadu. She has compiled A Badaga – English dictionary along with Paul Hockings (1992) which is applauded by scholars. Since Badaga has no script, the head entries are given in roman script.

Similarly, I along with Dr. S. Raja, completed a Tamil-Malayalam Translator's Dictionary. A Malayalam-Tamil Dictionary is also being compiled by us. The Centre for Applied Linguistics and Translation Studies, Hyderabad has a dictionary project. Under that they have taken up the works on Telugu-Hindi Nishantarvu, Telugu-Oriya nikhantuvu, Telugu-Kannada, Nighantuvu and Telugu-Malayalam Nighantuvu. Here the same strategy is employed. For one Telugu main entry there will be two or three target language equivalents, which are given in an order of frequency. I have been collaborating with the Telugu-Malayalam Nighantuvu. Prof. Umamaheswar Rao is

the Chief Editor for all the volumes. In addition to this he is engaged in projects concerning to online lexicography.

Tamil Lexicon was published by the University of Madras in 6 volumes. It has a supplementary volume also. Now Prof. Jeyadevan and the team in the University of Madras are engaged in the revision of the Tamil Lexicon by incorporating new entries. Similarly the compilation of Malayalam Lexicon in 11 volumes was entrusted to the University of Kerala, Trivandrum. 7 volumes have been published. The first volume was published in 1956. The works in connection with the volumes 8, 9 and 10 are going on simultaneously. In addition to this there are many good monolingual and bilingual dictionaries available in Malayalam. A Pedagogical dictionary in Malayalam (Sabda Surabhi 2 volumes, 2005) compiled by the former Malayalam Lexicon Chief Editor Dr. B C. Balakrishnan is noteworthy. In this he has incorporated all the Malayalam words available in the Malayalam books up to 12th Std. with details and meanings. This is a good lexical resource for future researches also. Dr. Shanavas of the Department of Linguistics, University of Kerala is engaged in the preparation of Malayalam corpus using modern technology

'Preparation of Generative Lexicon using MRDs' is an ongoing project undertaken by Prof. S. Rajdendran, Tamil University. Another project undertaken by him is 'Indian Languages to Indian Languages' (Machine Translation System). A bilingual mapping dictionary of Tamil-Malayalam and Malayalam-Tamil will be the byproduct of this project.

While talking about the 'Desiderata for a Modern English Dictionary of Tamil' Harold F. Schiffman opines that "The non-Tamils who learn an Indian language other than Sanskrit or Hindi is immediately aware of the problem of lack of adequate materials for learning the language, and especially the lack of decent dictionaries" (South Asian Language Review Vol.V, Jan. 1995. No.1).

Gregory James, the author of 'A History of Tamil Dictionaries' in one of his articles opines that "The nikantu remained the principal style of dictionary composition in Tamil from the earliest times until at least the 16th C.A.D". Tiva:karam with 9,500 entries 8th C.AD. is the first nikantu for Tamil language. Then pinkalam with 14,700 entries came into existence between 8th C and 13 C AD. Another popular nikantu called cu:ta:mani nikantu by Mantalapurutar, the first metrical lexicon was published during 1520 AD.

Dr. L. Shobha, Au-KBC Research Centre has undertaken a few projects the results of which will be useful for the creation of lexical resources for Indian languages. The ontological study is a must for information retrieval. Her work an 'Tamil-Hindi Transfer Lexicon' is useful in many ways. In the other important research project entitled 'Named entity' she has incorporated more than thirty thousand proper names of Tamil with the details. It seems this is the first attempt in any Indian language. An electronic thesaurus of Tamil has been prepared by the Department of Linguistics of Tamil University in collaboration with the Department of Computer Science of Tamil University. The CALTS, Hyderabad is engaged in the preparation of multilingual thesaurus It has to be noted that IIIT, Hyderabad is engaged in preparing lexical resources for Indian Languages. A team of scholars are involved in it.

Dr. T V. Geetha of Anna University of Chennai has done a good amount of research on the Tamil lexical resources under the Technical development of Indian Languages. Her online pedagogical dictionary (English-Tamil and Tamil-English) can be used easily by any body. As a whole the present scenario is bright on the preparation of lexical resources for Indian Languages.

V. WORDNET

"WordNet was originally conceived and developed as a lexical database for English on the basis of psycholinguistic properties. The major lexical categories like Nouns, verbs, adjectives and adverbs are organized in terms of sets of synonyms (synsets), each representing a lexical concept." (Dravidian WordNet, AU-KBC Research Centre, Chennai). The first version of Tamil WordNet has been prepared by AU-KBC and Tamil University with the financial support of Tamil Virtual University. It is going to be launched as an open source material in on-line

Developing WordNet is an important and basic need for lexical resources. The Euro WordNet developed for European languages is really useful for various kinds of research. Similarly if an India it will be a very good resource. The scholars who work on this field are aware of this phenomenon and hence, lot of such studies are in progress in India. It is to be mentioned that IIT, Mumbai has been engaged in the preparation of Hindi WordNet. It is available in the internet in its full-fledged form. This has to be appreciated since it is the first one from India.

I take this opportunity to appreciate the organizers for conducting this National Seminar on 'Creation of Lexical Resources for Indian Language computing and processing' with the clear cut objectives which are spelt out in the call letter itself An added attraction of this seminar, which is interdisciplinary in nature, is that linguists, computational linguistics, lexicographers, computational lexicographers, NLP/NLT researchers developers, computer scientists, etc. are involved in the deliberations. I am sure there will be a very good outcome in the field by the deliberations for all the three days.

ACKNOWLEDGEMENT

I thank Prof. K. Balasubramanian and Prof. S. Rajendran who helped me in the preparation of this paper.

REFERENCES

[1] Cruse, D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.
[2] Cuypers, I & G. Adraens. 1997. Periscope: the EWN Viewer. EuroWordNet Project LE4003. Deliverable D008d012. Amsterdam University of Amsterdam.
[3] Fellbaum, C. 1990. "English Verbs as a Semantic Net". International Journal of Lexicography, Vol. 3, No.4, 278-301.
[4] - - - - -. 1998. "A Semantic Network of English Verbs". In Fellbaum, C. (ed). WordNet. An Electronic Lexical Database. Cambridge: MIT Press.
[5] Fellbaum, C. (ed.) 1998. WordNet. An Electronic Lexical Database Cambridge, MA MIT Press
[6] Gross, D and K J Miller 1990 "Adjectives in Wordnet" International Journal of Lexicography, Vol. 3, No.4, 265-277.
[7] James, G. 2002. *colporul*: History of Tamil Lexicography. Chennai: CreA
[8] Louw, M. 1997. *The Polaris User Manual*. Internal Report, Lernout & Hauspie.
[9] Lyons, J. 1977. Semantics (Vol.1). Cambridge: Cambridge University press.
[10] Miller, G.A. 1990. "Nouns in WordNet: a lexical inheritance system". International Journal of Lexicography Vol 3, No. 4, 245-264.
[11] Miller, G A. 1991. Science of Words. New York: Scientific American Library.
[12] ----. 1998. "Nouns in WordNet". In: Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[13] Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database". International Journal of Lexicography, Vol. 3, No 4, 235-244.
[14] Miller, K.J. 1998. "Modifiers in WordNet". In: Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[15] Nida, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mounton.
[16] ----. 1975.b. Exploring Semantic Structure The Hague Mouton.
[17] Pustejovsky, J. 1995. The Generative Lexicon. Cambridge: MIT Press.
[18] Rajendran. S, 1978. Syntax and Semantics of Tamil Verbs. Ph D. Thesis. Poona: University of Poona.
[19] - - - -. 1983. Semantics of Tamil Vocabulary (Report of the UGC sponsored Postdoctoral Work in Manuscript) Poona: Deccan College Post-Doctoral Research Institute.
[20] - - - -. 1995. 'Towards a Compilation of a Thesaurus for Modern Tamil.' South Asian Language Review 5 1: 62-99.
[21] - - - - -. 2001. *taRkaalat tamizc coRkaLanjciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.
[22] - - - -. 2002. 'Preliminaries to the preparation of a Word Net for Tamil 'Language in India 2:1
[23] - - - -. 2003. 'Pre-requisite for the Preparation of an Electronic Thesaurus for a Tex Processor in Indian Languages'. Language in India 3:1.
[24] Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S Thiagarajan 2002 "Tamil WordNet " Proceedings fo the First International Global WordNet Conference Mysore CIIL, 271-274
[25] Rajasekharan Nair, N and S Raja 2006 A study of dialect variation in Tamil Agricultural vocabulary (mimeo).
[26] S. Sundarabalu. 2006. Semantic field and Lexical structure in Modern Tamil: Ph D. dissertation.
[27] Tengi, R.I. 1998. "Design and Implementation of the WordNet Lexical Database and Searching Software". In Fellbaum, C. (ed) 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.
[28] Uma Maheswar Rao, G. 2003. 'Some Issues in building Dravidian WordNet', Paper presented in the *Workshop on WordNet for Dravidian Languages*, Chennai: AU-KBC Research Centre, 2-3 June 2003.
[29] Vossen P. (eds.) 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
[30] Vossen, P. 2001. 'Condensed Meaning in EuroWordNet', In: P. Bouillon & F Busa (eds.). 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press, pp 363-383.
[31] Princeton English WordNet http: www.cogsci. princeton. edu.~wn
[32] EuroWordNet http: www.let.uva.nl ~ewn
[33] Global WordNet Association http: www. globalwordnet. org
[34] http framenet icsi berkeley.edu
[35] http wordnet.princeton edu
[36] http dictionary.cambridge.org

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.2 Study of Cognates among South Asian Languages for the Purpose of Building Lexical Resources

Anil Kumar Singh and Harshit Surana

Abstract— **High level of linguistic diversity in South Asia poses the challenge of building lexical resources across these languages. The only way we can hope to do this is by automating as much of this task as possible. This, in addition to the algorithmic aspect, also has a linguistic aspect in the sense that linguistic study can tell us what and how much can be automated. In this paper, we present a study of cognates across some South Asian languages for estimating how much of the task of building lexical resources can be automated. For identifying the cognates, we have used a unified computational model of scripts (UCMS) for Brahmi origin scripts. We have previously applied UCMS to solve several other practical problems. Based on the results of cognate identification, we suggest some implications for building lexical resources.**

Index Terms— **Cognates, South Asian languages, Building lexical resources, Unified Computational Model of Scripts.**

I. INTRODUCTION

CREATING large scale lexical resources is as difficult in terms of time and effort required, as it is important for building Natural Language Processing (NLP) applications or for linguistic reference. Given the number of major languages in India and the lack of financial and other resources, it may not be very practical to build lexical resources manually in the conventional ways. We need innovative ways to create such resources which can make use of computational power. However, as we discuss in the next section, automatically creating resources like bilingual or multilingual dictionaries has not been very successful so far. This is because the task is quite difficult To make this task easier, we need some linguistic insights. This is especially important in the South Asian context because South Asian languages have a lot of similarities [7] which can be abstracted out and used for computational purposes in solving problems which are otherwise not easy to solve.

In this paper we present a study of our experiments on automatic identification of cognates across some South Asian languages. We also suggest some implications of the result we obtained for building lexical resources

The method used for identification of cognates is based on a Unified Computational Model of Scripts [17] that we have previously used for solving several practical problems like spell checking, text normalization, improving information retrieval, shallow morphological analysis [18] etc.

II. SOME RELATED WORK

There has been some work on writing systems [25] from the computational point of view. Sproat [21] presented a computational theory of writing systems. He also studied Brahmi scripts [19] and even performed a formal computational analysis of Brahmi scripts [20]

Some other related work is on phonetic modelling of graphemes Rey et al. [12] argue that graphemes are perceptual reading units and can be considered the minimal 'functional bridges' in the mapping between orthography and phonology. Black et al. [1] have discussed some issues in building general letter to sound rules within the context of speech processing

Emeneau [7], in his classic paper 'India as a Linguistic Area', showed that there are a lot of similarities among Indian languages, even though they belong to different families

Our model of alphabet, which is a part of the UCMS, is based on the traditional knowledge about the scripts used for Indian languages and the work done on encodings for Indian languages. Perhaps the most important work in this category is the development of a standard for Brahmi origin scripts [4, 5], called Indian Standard Code for Information Interchange (ISCII). This has also been called a super-encoding or meta-encoding It took into account the similarities among the alphabets of Brahmi origin scripts.

Singh [16] had proposed a computational phonetic model of Brahmi based scripts based on orthographic and phonetic features. These features were defined based on the characteristics of the scripts. The similarity between two letters was calculated using an SDF and the algorithm used for 'aligning' two strings was dynamic time warping or DTW [11].

The unified model [17] also takes into account non-phonetic aspects of Brahmi scripts, like the aaksharik nature of these scripts and uses a very different way for calculating surface similarity.

The need for automatically extracting dictionaries has been recognized for a long time, which is natural since building dictionaries for various language pairs (especially in electronic machine readable form) is a long and difficult task for humans. There have been many attempts in this direction but the accuracies achieved so far have not been very high

Perhaps the biggest systematic effort at building multilingual dictionaries is the Papillon project [2]. It aims at 'creating a cooperative, free, permanent, web

Anil Kumar Singh (email: anil@research.iiit.net) and Harshit Surana (email: surana.h@gmail.com) are with Language Technologies Research Centre International Institute of Information Technology Hyderabad India

oriented environment for the development and the consultation of a multilingual lexical database'. But rather than separately creating dictionaries for different language pairs, it uses a set of monolingual dictionaries of word senses (lexies) linked through a central set of interlingual links (axies).

This linking of monolingual dictionaries can be done automatically [22]. Automatic extraction can also be done from comparable corpora [14]. Attempts have also been made to create WordNet-like lexical databases' [9]

An early attempt was Daelemans's [6] tool for automatic creation, extension and updating of lexical knowledge bases distinguishes between two levels of representation: a static storage level and a dynamic knowledge level.

Among others, Schiffman and McKeown [15] tried automatically building a lexicon of phrases from a collection of documents for question answering

Verma and Bhattacharyya [23, 24] have tried to automatic generate multilingual lexicon by using WordNet. Baud et al [3] describe a method to facilitate the interchange of lexical information for multiple languages in the medical domain. Farwell et al. [8] have used a method for automatic creation of lexical entries for a multilingual machine translation system.

Ribeiro et al. [13] have surveyed some of the algorithms for cognate alignment, including that by Melamed [10]. Their method is based on finding identical words as well as typical contiguous and non-contiguous character sequences extracted using a statistically sound method. Since they used this method for alignment of parallel text, we have not used it for comparison with our method as we are using non-parallel corpus.

III. SOUTH ASIA AS A LINGUISTIC AREA

South Asia has a common historical and cultural background. This is also reflected in the languages of this area. That most of these languages, in spite of belonging to various families, have a lot of similarity, was established formally in Emeneau's classic work [7]. From this work has emerged the idea of 'India as a linguistic area'. This occurrence is also called the South Asian convergence. So, in fact, it is better to talk about South Asia as a linguistic area.

There has been a debate about the reasons for this convergence, but whatever they might be, similarity among the South Asian languages is an established fact. This is understandable given the long term contact, migrations of populations, common historical and cultural background etc. It is as if South Asian languages form a family of their own, which cuts across different conventionally identified linguistic families.

IV. COGNATES AMONG SOUTH ASIAN LANGUAGES

Due to the 'convergence', South Asian languages have a lot of cognate words, i.e., their vocabularies have a significant overlap. For example, a lot of words in these languages have been borrowed from Sanskrit. Some of them retain their original form (tatsam), while others have changed their forms (tadbhav). Similarly, some of them have retained their meanings, while others have become associated with different concepts.

Apart from the large number of Sanskrit words, there are also a lot of words borrowed from languages which have been or still are dominant at some time in the history of South Asia. Persian and English are two such languages. Other 'foreign' languages like Arabic, Turkish and Portuguese etc. have also contributed their words to the South Asian languages, though in smaller numbers. Then there are the words which South Asian languages have borrowed from each other or from extinct or nearly extinct tribal and minority languages or dialects. Still another category of words which may be considered cognates are the onomatopoeic words

Identifying such cognate words can be a major step in calculating cross lingual lexical similarity in general. And, of course, taking care of cognates can help us in reducing the work involved in building lexical resources

V. SOME EXAMPLES

We can roughly divide the cognates into two categories. In the first category are those cognates which have more or less the same meaning in the two languages, while in the second the meanings have changed

Some examples of cognates extracted from our method are given below.

- **Same origin, same meaning :**
  - Hindi-Bengali : हाजार-হাজার, जीबन-জীবন, दोकानदार-দুকানদার, ओलट-ওলটা, सोहाग-সুহাগ, कालेक्टर-কলেক্টর etc.
- **Same origin, different meaning :**
  - Hindi-Bengali : One very good example of this is अभिमान This word is used in many Indian languages but has different meanings in different languages.

These two categories of cognates have to be handled differently for practical computational applications. While the first can be directly used for building lexical resources, the second cannot be used directly.

VI. LEXICAL SIMILARITY AT THE SURFACE

Surface similarity can be divided into two overlapping parts: orthographic and phonetic. For calculating such similarity, we have used a unified model of scripts [17]. This model is very useful for South Asian languages because of the phonetic and aaksharik (loosely, syllabic) nature of Brahmi origin scripts

Earlier work on connecting phonology and orthography has focused on letter to phoneme (or vice-versa) mapping. We have used a model of scripts based on phonetic and orthographic features of letters, a stepped distance function (SDF), the aaksharik nature of the scripts and a shallow model of morphology, etc. We can use either the Computational Phonetic Model of Scripts (see that next section) alone or we can use the UCMS. The phonetic and orthographic features are manually defined based on the characteristics of the scripts. In the CPMS, similarity of two strings can be calculated by using a dynamic programming algorithm called the dynamic time warping (DTW) algorithm.

VII. UNIFIED COMPUTATIONAL MODEL OF SCRIPTS

The UCMS [17, 18] aims to capture the characteristics and commonalities of a group of scripts, restricted to the Brahmi origin scripts for the time being The idea is that instead of only calculating literal string similarities or even phonetic similarities, we can use all the information that would become available to the computer if the computer knew about the linguistic characteristics of the scripts. This would make out applications more accurate and more flexible, and these applications might work well across all the languages which use the scripts covered by the model

The schematic diagram of the UCMS is shown in figure-1.

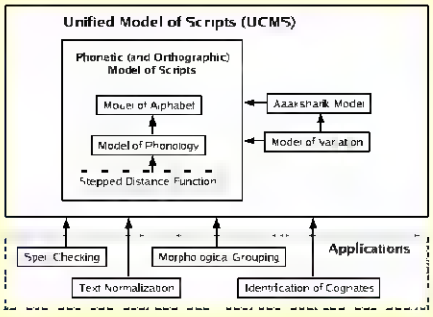


Fig.-1: Unified Computational Model of Scripts

As can be seen from the figure, the unified model consists of various component models. The most important part of the UCMS (for Brahmi scripts) is the Computational Phonetic Model of Scripts (CPMS). The CPMS is itself composed of a model of alphabet and a model of phonology, a stepped distance function for calculating the similarity of letters or akshars, an alignment algorithm for calculating the similarity of words or strings. Other parts of the UCMS are a model of variation and a shallow model of morphology (Singh and Surana, 2007a). The latter has not been shown in the figure. Some of the applications for which we have already used the UCMS are shown in the figure.

VIII. AUTOMATICALLY IDENTIFYING COGNATES

The cognate identification or extraction algorithm was based on using the UCMS. Surface similarity scores were calculated for pairs of words from different languages and a threshold was applied. The search was fast enough because we are using an FST of akshars. A dynamic programming based algorithm was used for aligning strings on the FST. Note that this algorithm is different from the one used in CPMS The CPMS was used for calculating akshar pair similarity scores. Since these scores were used while aligning the strings, these calculations were performed only once. The list of possible akshars was extracted from the corpus

For our experiments, we have used the ERDC parallel corpus. The cognate extractions algorithm was run on word lists (along with frequencies) extracted from the corpus. Only the top 20000 or so words were used for identifying cognates.

IX. COGNATE IDENTIFICATION EVALUATION

Since it was not possible to extensively prepare manual reference data across all the language pairs, we used random sampling to evaluate cognate extraction. We extracted cognates by applying a threshold and then randomly extracted 200 candidate cognates from them These lists of 200 cognates were manually checked by people who knew the two relevant languages quite well

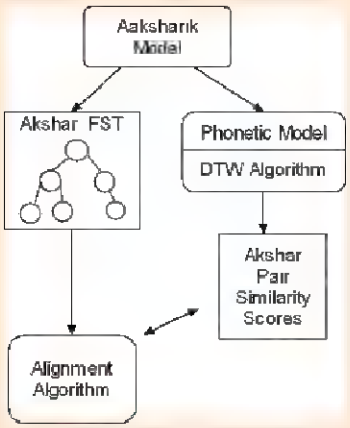


Figure 1: Identifying Cognates Using the UCMS

X. ESTIMATES OF COGNATE COVERAGE

By coverage of cognate coverage we mean what percentage of words in documents in South Asian languages are cognates for particular pairs of languages. Table-1 gives rough estimates of the coverage of cognate among several language pairs. Right now we have not calculated the coverage of cognates across more than two languages.

The estimates are only rough estimates because we have extrapolated from the results of our evaluation of the cognate identification algorithm and these estimates might include proper nouns (around 5% of extracted cognates). The validity of these estimates may also be different for different language pair because the algorithms used for cognate identification are, naturally, asymmetric. However, to counter this problem, we tried to make the algorithm as symmetric as possible. This means that no language-pair specific tuning of the algorithm was performed for the purpose of calculating cognate coverage, even though we could have used such tuning to improve the accuracy of identification for certain pairs of languages.

Since the size of the ERDC corpus for different languages was different, we compensated by performing the calculations only for a particular range of frequencies. The token list size for each language was more than 20000 words

XI. IMPLICATIONS FOR BUILDING LEXICAL RESOURCES

One inference of the results of cognate identification that is clear from table-1 is that the percentage of cognates that can be extracted automatically depends on the similarity (or distance) between the two languages. For example, Hindi and Punjabi or Hindi and Bengali are much closer than Hindi and Telugu or Telugu and Marathi. Accordingly, many more cognates can be extracted for the first two pairs than for the last two. This is along the expected lines. Still, the results give us a quantitative estimate of the coverage of cognates. However, it is important to note that these results are from a particular corpus and are valid more for written language. Spoken language is likely to have a different distribution of cognates, e.g., there will be fewer Sanskrit origin tatsam words in spoken language, but more words of English or Persian origin. Another important point is that these figures are for tokens, not types. Since we are mostly interested in written language, these results can still be useful for us.

Hindi-Punjabi	**57.63**
Hindi-Marathi	**44.54**
Hindi-Telugu	**28.86**
Hindi-Bengali	**50.43**
Telugu-Marathi	**28.62**
Telugu-Kannada	**34.67**
Bengali-Assamese	**55.48**

Table-1: Percentage Cognates (Tokens)

Another implication from our observation of the output is that some language pairs like Hindi-Telugu have cognates mostly in the category of tatsam words, whereas pairs like Hindi-Punjabi have cognates of other kinds too. This means that the task of preparing practically useful lexical resources is even more difficult than what appears from the figures given in table-1.

The results also show that if surface similarity is calculated in a more linguistically aware way, then just by calculating such similarity we can get a lot of crosslingual information that can be used for building lexical resources

A surprising result is that Telugu and Kannada have significantly fewer cognates than Hindi-Marathi, though the first two are linguistically and geographically supposed to be closer than the last two. If our figures are valid, and not heavily biased by the corpus, then it implies that the ease of building lexical resources may not be in direct proportion to the linguistic distance between two languages.

For building multilingual lexicon, our results show that we have to rely on more intelligent applications for extracting corresponding words across languages. Such applications have to take into account the characteristics of languages or language pairs. An algorithm that has no way to incorporate linguistic knowledge will not work equally well for all languages or language pairs. The challenge is to design applications which can do this with minimum human intervention

More specifically, as mentioned earlier, the two categories cognates have to be handled differently. Since the second kind of cognates, i.e. words which occur across different languages but with different meanings, we need to combine the linguistics knowledge based approaches with machine learning based approaches. We could distinguish between the two kinds of cognates and the meanings of the second kind by using a method based on contextual similarity, somewhat like the methods used for word sense disambiguation.

XII. CONCLUSION

In this paper we have presented a study of the results of automatic cognate extraction from non-parallel multilingual corpora of some South Asian

languages using a Unified Computational Model of Scripts, which we have previously applied to several other practical problems. We argue that these results have some implications for building lexical resources, both from linguistic and computational points of view. Taking care of these implications might make the task of building lexical resources easier, given the scarcity of financial and other resources for this task and the high linguistic diversity in the South Asian region.

REFERENCES

[1] Black, A., Lenzo, K.; and Pagel, V. 1998. Issues in building general letter to sound rules. In ESCA Synthesis Workshop, Australia , 164171.

[2] Boitet, C., Mathieu Mangeot-Lerebours, M and Gilles, S. 2002. The PAPILLON Project: Cooperatively Building a Multilingual Lexical Database to Derive Open Source dictionaries and Lexicons. Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop Taipei, Taiwan

[3] Baud, R H., Nystram, M , Bonn, L. Evans, R., Schulz, S. and P Zweigenbaum. 2005. Interchanging Lexical Information for a Multilingual Dictionary

[4] Bureau of Indian Standards. 1991. Indian standard code for information interchange (ISCII)

[5] C-DAC. 2006. Standards for Indian languages in it. http www cdac in/html gist standard asp

[6] Daelemans, Walter. 1987. A tool for the automatic creation, extension and updating of lexical knowledge bases. Proceedings of the third conference on European chapter of the Association for Computational Linguistics. pp 70-74. Association for Computational Linguistics.

[7] Emeneau, M. B. 1956. India as a linguistic area. In Linguistics 32:3-16

[8] Farwell, D , Guthrie, L. and Wilks, Y. The Automatic Creation of Lexical Entries for a Multilingual MT System In Proc. of COLING 2002

[9] Lipatov, A., Goncharuk, A., Helfenbein, I , Shilo V. and Lehelt, V. 20003. Automatic Creation of Non-English WordNet-like Lexical Databases. Papillon 2003 Workshop. NII, Tokyo, Japan

[10] Melamed, I. D. 1997. A portable algorithm for mapping bitext correspondence. In Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 305312. Association for Computational Linguistics

[11] Myers, C. S., and Rabiner, L. R. 1981. A comparative study of several dynamic time-warping algorithms for connected word recognition In The Bell System Technical Journal, 60(7), 13891409

[12] Rey, A.; Zieglerc, J. C.; and Jacobse, A. M. 2000 Graphemes are perceptual reading units. In Cognition 74

[13] Ribeiro, A.; Dias, G.; Lopes, G.; and Mexia, J. 2001. Cognates alignment. Machine Translation Summit VIII, Machine Translation in The Information Age 287292

[14] Sadat, F. 2002 A Combination of Models for Bilingual Lexicon Extraction from Comparable Corpora. Papillon 2002 Seminar. NII, Tokyo, Japan

[15] Schiffman, B. and McKeown, K. 2000 Experiments in automated lexicon building for text searching. In Proc. of COLING-2000

[16] Singh, A. K. 2006. A computational phonetic model for indian language scripts. Constraints on Spelling Changes: Fifth International Workshop on Writing Systems Nijmegen, The Netherlands

[17] Singh, A. K. and Surana, H 2007a. There can be Depth in the Surface: A Unified Computational Model of Scripts and Its Applications. Under submission.

[18] Singh A.K. and Surana, H. 2007b. Using a Model of Scripts for Shallow Morphological Analysis Given an Unannotated Corpus. ADD-2 Morpho-Syntactic Analysis (2nd School of Asian ANLP for Linguistics Diversity and Language Resource Development) , Pathum Thani, Thailand

[19] Sproat, R. 2002. Brahmi scripts. In Constraints on Spelling Changes: Fifth International Workshop on Writing Systems

[20] Sproat, R. 2003. A formal computational analysis of indic scripts. In International Symposium on Indic Scripts. Past and Future

[21] Sproat, R. 2004. A Computational Theory of Writing Systems

[22] Teeraparbseree, A. 2003. Jemime: A Flexible System for the Automatic Creation of Interlingual Databases. Papillon 2003 Workshop NII, Tokyo, Japan

[23] Verma, N. and Bhattacharyya, P. 2003. Automatic Generation of Multilingual Lexicon by Using WordNet. International Conference on Convergence of Knowledge, Culture, Language and Information Technology. Library of Alexandria, Egypt

[24] Verma, N. and Bhattacharyya, P. 2004. Automatic Lexicon Generation through WordNet. Global WordNet Conference (GWC-2004), Czech Republic

[25] Coulmas, Florian Writing Systems: An Introduction to their Linguistic Analysis. Cambridge University Press, 2003

This paper was presented at LRII-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt of India.

8.3 Bricks and Mortar for Digital Resources in Indian Languages

Gora Mohanty

Abstract **The creation of resources for building Indian language content in the online world needs various kinds of support technology. Over the years, free software tools have matured enough to provide a stable and easily usable base for supporting Indian languages in a standardized, platform- neutral manner. I will present a complete set of such tools that can immediately meet needs for content generation in most major Indian languages.The talk will cover text input methods, spellcheckers, sorting and locale issues, font converters, transliterator, databases, and text-to-speech, and focus on a live demonstration of such tools. I will also emphasize the importance of free software, and of open document formats in this area.**

I. INTRODUCTION

TILL recently, computers have by and large been easily usable only in English, and, to some extent, other European languages. What Indian language software has been available has usually operated in a non-standardised manner, so that severe issues have arisen with the portable use of content created with such tools Things have changed over the last few years, with improved support for at least the major Indian languages, and with the advent of user interfaces in native Indian languages.

Open-source operating systems, and applications, have traditionally held a lead in supporting multi-lingual features, and support for international standards like Unicode [29], and for the complex text handling required by Indian languages has also matured. This article discusses the open-source tools available for content creation and management in Indian languages, covering both minimal requirements like fonts, input methods, and rendering, as well as advanced tools like spell-checkers, and text-to-speech. It also explains the importance of free software in this context, and outlines possible ways of spurring the creation of online content.

II. FREE SOFTWARE IN THIS CONTEXT

Open-source software provides four basic rights to the end-user, namely the right to use, study, modify, and redistribute the modifications, whereas closed-source software provides only the right of use. These rights have some far-reaching consequences. The ability to study the software lets new developers learn programming techniques from the world's top experts. The right to modify, and redistribute the software ensures that one is not left at the mercy of any single organisation. The net result is to commoditise software, ensuring that it remains economical. As it is not feasible in an open-source environment to continuously milk consumers by charging an arm and a leg for each new version of an application, the commercial focus of companies changes to providing high-quality support

In the context of localisation, open-source software complies better with standards, and, in particular, there is system-wide compliance with Unicode. There is also a built-in internationalisation framework that allows easy translation to local languages, so that almost every open-source application nowadays permits localisation. These advantages have been recognised even by hard-nosed economists, e.g., the conservative Economist magazine [13] noted back in 2003 that open-source desktop interfaces were available in more than twice as many languages as Windows XP, a lead that has since widened. Licensing costs for open-source software are also typically zero, which is a huge benefit for disadvantaged portions of society, as well as for non-profit institutions such as educational organisations, NGOs, and government bodies.

Finally, one of the interesting aspects of open-source software development is that it encourages a community-based effort, as the end-user now can, and in fact is actively encouraged to participate in the development process, rather than being a passive consumer. This spirit of sharing, and joint effort, is important if one seeks a bottoms-up, rather than a top-down approach. It is not enough to simply hand out free-of-cost copies of software in the name of bridging the digital divide. One has to ensure that the linguistic community that is supposedly being served by such software becomes an integral part of the process, and actually make use of the basic tools to create things of lasting value

III. ESSENTIAL SOFTWARE

There are four indispensable items for an Indian language to be supported by an operating system:

- System-level support for Unicode, and rendering Unicode support has been available on Linux for a while, and extends to all levels of the support from the base C libraries, glibc, to X and various rendering engines. The complex features required for the rendering of Indian language text are usually provided through font technologies like OpenType [2, 10, 9], developed jointly by Adobe and Microsoft, and also supported on newer Apple computers. OpenType for Indic languages is supported in Linux through various renderers like ICU [8], IndiX [3], Pango [23], and QT [25], though it should really be integrated into the base X GUI layer.

Gora Mohanty is with Sarai CSDS, 29 Rajpur Road New Delhi (e-mail. gora@sarai net)

- Fonts An OpenType font, with Unicode encoding, is required for rendering, printing, etc. A variety of such fonts, covering most major Indian languages, are available under an open-source licence, e.g., such as the one listed under the IndLinux project [1].

- Input methods: Here, we confine our attention to keyboard input methods, though in the long run, other forms of input, such as handwriting recognition, or speech-to-text might be more useful in an Indian context. Rather than making custom keyboards for each Indian language, it makes more sense to simply map characters in existing English keyboards to Indian language equivalents through software. Under Linux, this mapping can be done through various applications, but we will focus on only two of these: (a) xkb [5, 31, 7] provides a very low-level mapping of the keyboard, interfacing directly with the base X GUI layer, and (b) SCIM [26] that provides many more possibilities than the simple one-to-one mapping scheme of xkb. SCIM is probably the future of open-source keyboard input methods.

  There are two broad classes of keyboard maps in use for Indian languages: (a) pseudo-phonetic ones like ITRANS [4], Bolnagri (a xkb-based phonetic map), etc., and (b) non-phonetic ones like Inscript, and Remington (typewriter) keymaps, which aim at efficiency in typing.

- Locales: A locale defies culturally-sensitive information, including things like names of months, days of the week, currency symbols, sorting order, etc. Locales are usually system wide, such as the locales for various Indian languages that come bundled with glibc, though some applications, like OpenOffice, prefer to use their own.

The support for these items need to be available at a very low-level in the system, so that applications share a common code-base, making it easier to resolve any bugs and deficiencies in the software. The old practice of each application rolling its own support for Indian language features leads to a myriad issues with non-standardisation, and incompatibility with other software.

Position	Hindi	Default Eng.	Best Eng.
Not found	5%	6%	2%
1	71%	59%	60%
1-5	91%	86%	83%
1-10	94%	91%	90%
Any	95%	94%	98%

Table 1: Performance of aspell in spell-checking Hindi, and English.

In addition to these, the following items are also important for a modicum of comfort in working in Indian languages on the computer:

- Spell-checking: The GNU spell-checking engine, aspell [18] now includes support for documents in UTF-8 (Unicode). Dictionaries for many Indian languages are also available [6], though they are far from comprehensive at the moment. aspell has many nice features, such as phonetic rules and affix specifications, that make it well-suited for spell-checking in Indian languages.

  We have recently gone through the exercise of customising aspell for Hindi, with the aim of also using this as a case study for other Indian languages. Table 1 compares the performance of aspell in Hindi against the default English method, and also against the best procedure for English (this is not used by default as it is more time-consuming). A test list of deliberately mis-spelled words was run through aspell for each language, and the performance was measured by (a) counting the percentage of words for which the correct replacement was suggested, and (b) the position of the correct replacement in the list of suggestions returned by aspell. It can be seen from the table that the performance for Hindi typically exceeds that for English. This is not as surprising as it might be at first glance, as Hindi is spelled phonetically.

- Sorting: A default sorting order for Indian languages has been defined in the glibc locales, but has been tested only for Hindi, and Oriya. A more rigorous testing should be undertaken, and incorporated into the Unicode Common Locale Data Repository [28].

- Printing: Printing in Indian languages now works from GNOME/KDE applications, as well as from browsers like Mozilla, and Firefox.

- Desktop software: Almost any desktop application required for homes, and businesses can also be used in Indian languages, under Linux. This includes office suites like OpenOffice [11], browsers like Firefox [17], mail clients like Zimbra [30] or Mozilla Thunderbird [21], databases like PostgreSQL [24] or Mysql [22] with front-ends like oobase, the OpenOffice database component, or rekall [12].

- Font converters, for content in legacy 8-bit fonts: Before the advent of standardised encodings like ISCII or Unicode, software for creating Indian language content used non-standard, 8-bit font encodings, so that there is a large body of existing content in these formats, making it useful to have a means of conversion to Unicode. There is a Firefox plugin, Padma [20], that does on-the-fly conversion of online content from these legacy encodings to Unicode, and work is under way to have a general-purpose library, and utilities, to handle such conversion tasks.

- Dictionaries / glossaries: aspell includes dictionaries for several Indian languages, but these are far from being large enough, and comprehensive enough. Likewise, glossaries of technical terms have been prepared for use in translations of GNOME, KDE, OpenOffice, etc., but a lot more work is needed here

IV. ADVANCED SOFTWARE

Besides the basic tools listed in the previous section, work is at various stages of progress on advanced software that can help in Indian language content creation, and dissemination:

- Text-to-speech: Open-source software is now available that can convert electronic text into speech. This is useful, for example, to allow visually handicapped, or illiterate, people access to online content. There are two main open-source applications that can synthesise speech in Hindi, and a few other Indian languages: (a) Festival [16] is the better-known system, which has modules for Hindi, Marathi, and Telugu, and (b) eSpeak [15] which uses a different technique to synthesise speech. For both applications, there remains a fair amount of work in getting the Hindi speech synthesis to work accurately
- Speech-to-text: The opposite problem, of the recognition of speech, and its conversion into electronic text, is a significantly more difficult. Among the various open-source speech recognition programs, perhaps the best-known is CMU Sphinx [14]
- Optical character recognition (OCR): This is another area which needs a lot of work, especially in the open-source domain. Of the open-source OCR engines, perhaps the most promising ones are GOCR, JOCR [19], and Tesseract [27].

V. PROCESS OF CONTENT DEVELOPMENT

The applications described here constitute only the essential tools, based upon which the real work of content development in Indian languages can be taken up. The modalities of how the development of meaningful content in Indian language will come out remain to be seen, and, in an open-source world, will eventually be driven by the users of the system themselves, but here are some ideas on how to spur the growth of online content in Indian languages:

- Change focus from technological development to deployment of existing solutions, and promotion of usage: The current set of tools in the open-source world are already good enough for creating Indian language content. The ongoing work of translation of the user interface of the applications into the local language will render them suitable for use even by non-English speakers. Thus, it is time now to put these tools in the hands of actual users, thereby empowering them to create content on their own
- In the immediate term, the quickest way to build a community of local language users is by building sites that cater to specialised interests. This could include items like blogs, news and search engines that work in Indian languages, and provide facilities like email to users. A good example of this is http://sampada.net which started out as a Kannada localisation site, but has slowly transformed into a focal point for literary writers, and other people interested in Kannada literature
- The other immediate area that can be targeted is the provision of these tools to vernacular educational institutions, and establishments that publish in Indian languages, people that already have a need for software in Indian languages, and who are already in the business of content creation.
- Wikipedia (http://en.wikipedia.org) is a very interesting project to create a free encyclopedia based solely on contributions from users. It is founded on principles similar to those of the free software movement, and has been a resounding success, being comparable in quality to top-of-the-line commercial encyclopedias like the Encyclopedia Britannica. Indian language versions of Wikipedia also exist, for example, the Hindi one (hi.wikipedia.org). Thus, the building of an active community of Indian language contributors to such an enterprise would be highly beneficial

REFERENCES

[1] A list of Indic fonts, with Unicode encoding
http://indlinux.org/wiki/index.php/IndicFontsList

[2] Adobe Open Type fontpage
http://www.adobe.com/type/opentype/main.html

[3] An Indian language compilation of GNU/Linux software
http://www.cdacmumbai.in/projects/indix/.

[4] An Indian language transliteration package
http://www.aczoom.com/itrans

[5] Description of the X keyboard extension
http://netadmin1.ic.tsu.ru/en/xkb/.

[6] Dictionaries for GNU Aspell
http://aspell.net/XXX

[7] Enhancements to xkb configuration.
http://www.xfree86.org/current/XKB-Enhancing.html

[8] ICU is a widely used set of C/C++ and Java libraries for Unicode support, software internationalization and globalization

http://www.306.ibm.com/software/globalization/icu/index.jsp. The Sourceforge ICU project is hosted at http://icu.sourceforge.net/.

[9] Microsoft FAQ on OpenType.
http www.microsoft.com typography faq faq9 htm.

[10] Microsoft OpenType font specifications.
http www.microsoft.com typography specs default.htm

[11] OpenOffice.org is an open-source, multi-platform and multi-lingual office suite, compatible with all other major office suites, and is free to download,use, and distribute.http www openoffice.org

[12] Rekall, the database management system. http: www.thekompany.com products rekall .

[13] Open source's local heroes. Multi-lingual support in open source software. *The Economist*, Dec. 4 2003. Online at http: www.economist.com science tq displaystory cfm?story id=2246308. Subscription might be required.

[14] The CMU Sphinx Group Open Source Speech Recognition Engines. http: cmusphinx.source forge.net/html cmusphinx.php.

[15] The eSpeak open source software speech synthesizer. http: espeak.sourceforge.net. eSpeak uses a different approach to synthesising speech than other open-source text-to-speech engines. A Hindi module is in the process of refinement.

[16] The Festival speech synthesis system.
http www cstr ed ac uk projects festival
A Hindi module for Festival can be obtained from
http: janabhaaratii.org.in:9673 indicbhaaratii/Members Pritu Patil festival-hi-0-1-tar gz.

[17] The FireFox browser homepage. http: www.mozilla. org products firefox

[18] The GNU Aspell homepage.
http aspell net

[19] The GOCR program, also known as JOCR
http: jocr.sourceforge.net .

[20] The homepage of the Padma Firefox extension.
http padma mozdev org
Padma does on-the-fly conversion from 8-bit Indian language fonts to Unicode.

[21] The Mozilla Thunderbird mail client
http www.mozilla.com en-US thunderbird .

[22] The Mysql database.
http mysql org

[23] The Pango project.
http www pango org

[24] The PostgreSQL database.
http: www.postgresql.org.

[25] The Qt cross-platform application development framework.
http: www.trolltech.com products qt.

[26] The Smart Common Input Method (SCIM) platform project.
http: www.scim-im.org .
Provides a user-friendly input method interface for POSIX-style operating systems, including a platform to make input method development easier

[27] The Tesseract optical character recognition engine.
http: code.google.com p tesseract-ocr .
This is based on code that was developed at HP Labs between 1985 and 1995, and has now been open-sourced. Most of the current work on Tesseract is sponsored by Google.

[28] The Unicode Common Locale Data Repository (CLDR).
http www unicode org XXX

[29] The Unicode Consortium
http: www.unicode.org

[30] The Zimbra mail server.
http: www.zimbra com.

[31] K. Toman and I.U. Pascal.
The xkb configuration guide.
http: www.xfree86.org current XKB-Config.html

This paper was presented at LRII-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.4 Evolving Translations & Terminology - The Open Source Way

G. Karunakar and Ravishankar Shrivastava, IndLinux.org

Abstract—**Localization is the process of enabling a product to suit a different cultural locale, in terms of information processing, language expression, and interface. FOSS model applied to localization has brought about a rapid enablement of several useful softwares to a wide array of languages.**

Index Terms—**Localization, Free and Open Source Software (FOSS), translations, terminology.**

I. INTRODUCTION

TRANSLATION has traditionally been implied with translating literary text, publicity information, official documents, books, news, communication, technology, etc. While in the former modes the vocabulary was sufficient or could represent new objects and subjects, there has been lack of it for technology evolving at a fast pace. Traditional language resources have not had the vocabulary to be used with the newer inventions & discoveries happening day by day. Post independence there were efforts to compile and define terminology dictionaries, while being comprehensive covering many areas of modern science, technology & society, they fell short of being widely accepted.

With the advent of computers and IT revolution, there has been a new set of vocabulary getting in common usage, which has kind of gone untouched by translation process until recently. Software interfaces have traditionally been English based because their evolution has primarily been in English-speaking locales. Though after the availability of more advanced programming techniques and a increasingly commercial market for software products forming globally there has been a rise in multilingual interfaces to software, which have primarily been brought by market demand or by government dictum (that software sold in a specific state/country be available in the national language).

Unavailability of software in local language has created a kind of digital divide. But an important aspect of the divide being that suitable terminology around the new technologies has not evolved at the same pace as the technology itself. And with respect to Indian languages the problem has been more acute since there has not been much of early concerted effort to evolve terminology for software interfaces in IT domain. There have been independent efforts in evolving terminology but not in a inclusive way to set a standard for it. Even if available they have failed to get wide acceptance due to different reasons.

So translating IT terminology into Indian languages has been left to the market players, who have made their own versions often differing in the translation and coinage of words for new terms. An example being Indian language interfaces of proprietary products like Windows XP, Lotus notes and other Indian language tools available from different commercial vendors.

New in this domain has been Free & Open source software which by its inherent philosophy allows easy modification by anyone interested. Since the source code is available freely (as in allowing modification and re-distribution), open-source software has been translated to many languages in the world, and also available in many Indian languages. This paper will outline the learnings of open-source based translation efforts.

II. THE FREE & OPEN SOURCE MODEL

The FOSS model is different in its functioning than a traditional development model, in that it's more open in its scope and operation than a closed source. While not going into technical details, the important perspective offered by FOSS model is its collaborative ways of working, peer review and philosophy of sharing. Internet plays the role of a backbone and glue to make it all possible.

In brief the life-cycle of FOSS project is like below:

- Project begins by individual effort/ motivation
- Initial version released in open license where source code is available for anyone to try and use.
- More people get involved starting as users & then as contributors since they find it useful & interesting
- With more contributions & testing the project matures.

G Karunakar is with Sarai, CSDS, 29 Rajpur Road, New Delhi (e-mail. karunakar@indlinux org)

Ravishankar Shrivastava is based in Ratlam, Madhya Pradesh (e-mail. raviratlami@gmail com, blog. http. raviratlami blogspot com)

- In time reaches a point where it can sustain itself, with the significant no of contributors
- The open & inclusive model means anyone with requisite skills can join to contribute in different ways-
  - like writing source code
  - testing
  - writing documentation
  - advocacy
  - translations

An important aspect of FOSS model is that the contributors come from varied geography, with different cultural backgrounds, languages, so the obvious fallout of it being that the FOSS software has multilingual capabilities by design. And the localization process is simple enough for anyone to contribute to translation efforts. Our interest is in the last part that is about how translation or creating lexical resources can benefit from a FOSS model.

III. FOSS TRANSLATION MODEL

FOSS localization has a good working model based on message catalogs, where all translatable strings of a software are extracted into a catalog, more like a database, which is then translated into multiple language with per language catalogs.

Typically there is a translation team for a language, which coordinates translation effort for that language. They keep a track of translatable catalogs coming in & complete translations based on work already done before. Tools available provide for rough translation and tracking changes across catalogs.

A. Message catalogs

Message catalog, also called Portable Object (PO), contains string pairs of original English string and its translation. An untranslated pair will have an empty translation entry. String duplication is avoided by having a common string pair for all similar occurrences. Across multiple files translations can be reused, i.e. if terms like "File", "Edit", "Save as" etc occur in multiple catalogs, an existing translation can be reused, rather than creating a new one, in a way keeping consistency. Catalogs can be updated for changes in strings & translations. A PO file also contains a header for meta-information like last translator, file creation date, updation date, language, etc.

B. Processes

The translation process starts by the formation of a language team. Typically this is driven by a loose team of motivated individuals. The team starts translation by taking up essential parts of a software, typically libraries, which are used by other application components. If it is a new language with no existing translation history or any standardized glossary available, then it has to start the painstaking effort of setting up a standard glossary by a constant process of review and learning. The first version of translations may not be the best, but gives a taste of efforts needed apart from exposing the problems in terminology evolution. The successive revision and translations can take this experience into account and improve upon that. Since the translations are also publicly available, it is easy to get comments on terminology and feedback from users. The same can then be incorporated into future activity. This cycle of feedback, learning and innovation ensures that over the time once the efforts reach a maturity level (say by working for 1-2years or completing a big volume of translations) a standard is reached in the translation terminology and quality is consistent.

IV. TOOLS

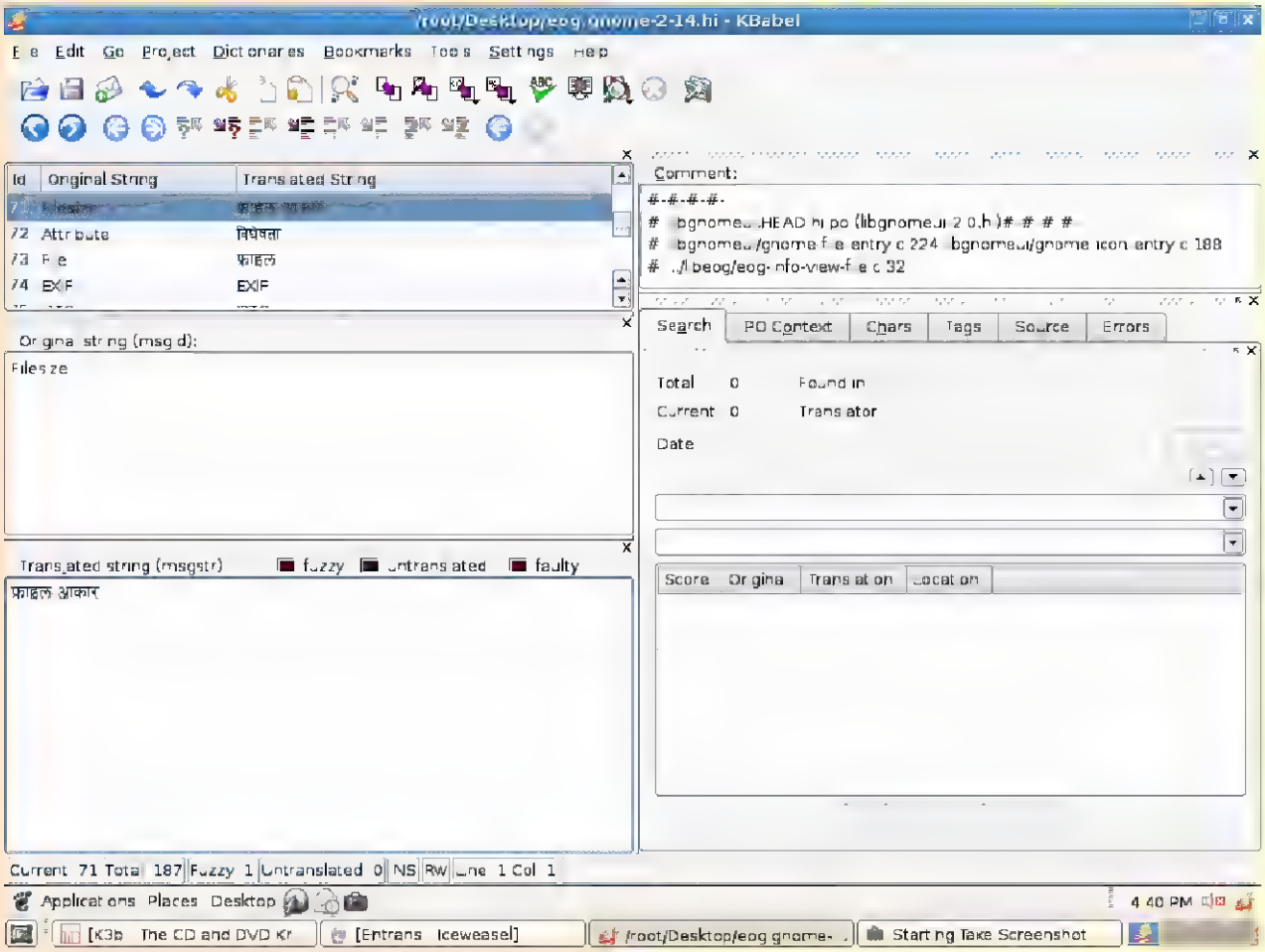
The tools make all the difference. Tools allow for collaborative working and reusing existing work. Broadly the tools can be classified as:

- Offline where translator works on own computer not connected to internet, updates are local.
- Online where multiple translators can work on common pool of strings while being connected to internet or over a local network.
- Communication tools like email, instant messaging, forums and mailing list are available for discussing translation issues & distributing & coordinating work among team members.

A. Kbabel

It is the most popular & powerful tool used in translations. It supports managing multiple catalogs. A simple and fast interface gives facility to add translations for new strings with support for rough translations based on existing translations. It can build a database of existing translations to give suggestions for the

untranslated entries, which the translator can choose to use. It also checks for common technical errors in translations, apart from giving facility to integrate with a spell checker. While the tool itself is used at individual level, duplication of work and consistency is achieved by coordinating distribution of work and sharing of translation memory (also called PO compendium) among the translation team.

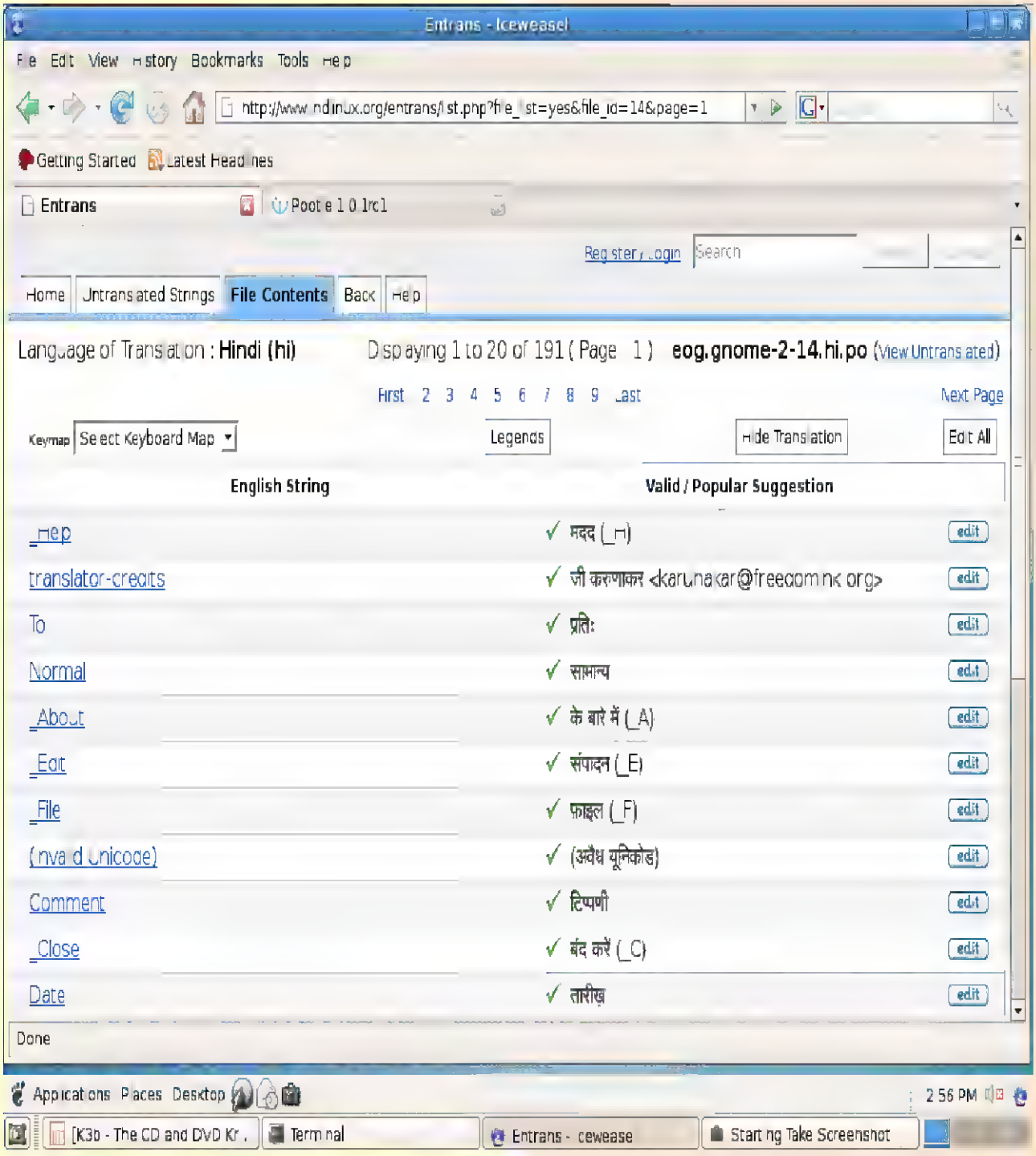


B. Entrans

An online translation tool, where message catalogs can be uploaded onto the server and a common pool of strings formed, which still keeps them grouped based on their source. It provides for the following:

- Registration for translators & validators.
- Anonymous users visiting site can go through strings and suggest translations and also vote for suggestions.
- Rough translations automatically get generated based on existing translations, which are available as machine suggestions.
- Translators can add new translations, which get included as suggestions.
- Translators, given the role of a validator, can select from multiple suggestions and approve translations.
- While translating/reviewing a string the interface shows multiple options, of which, one can be selected and approved or a different one can beprovided.
- A quick lookup search is provided for looking up translations for common words and phrases.

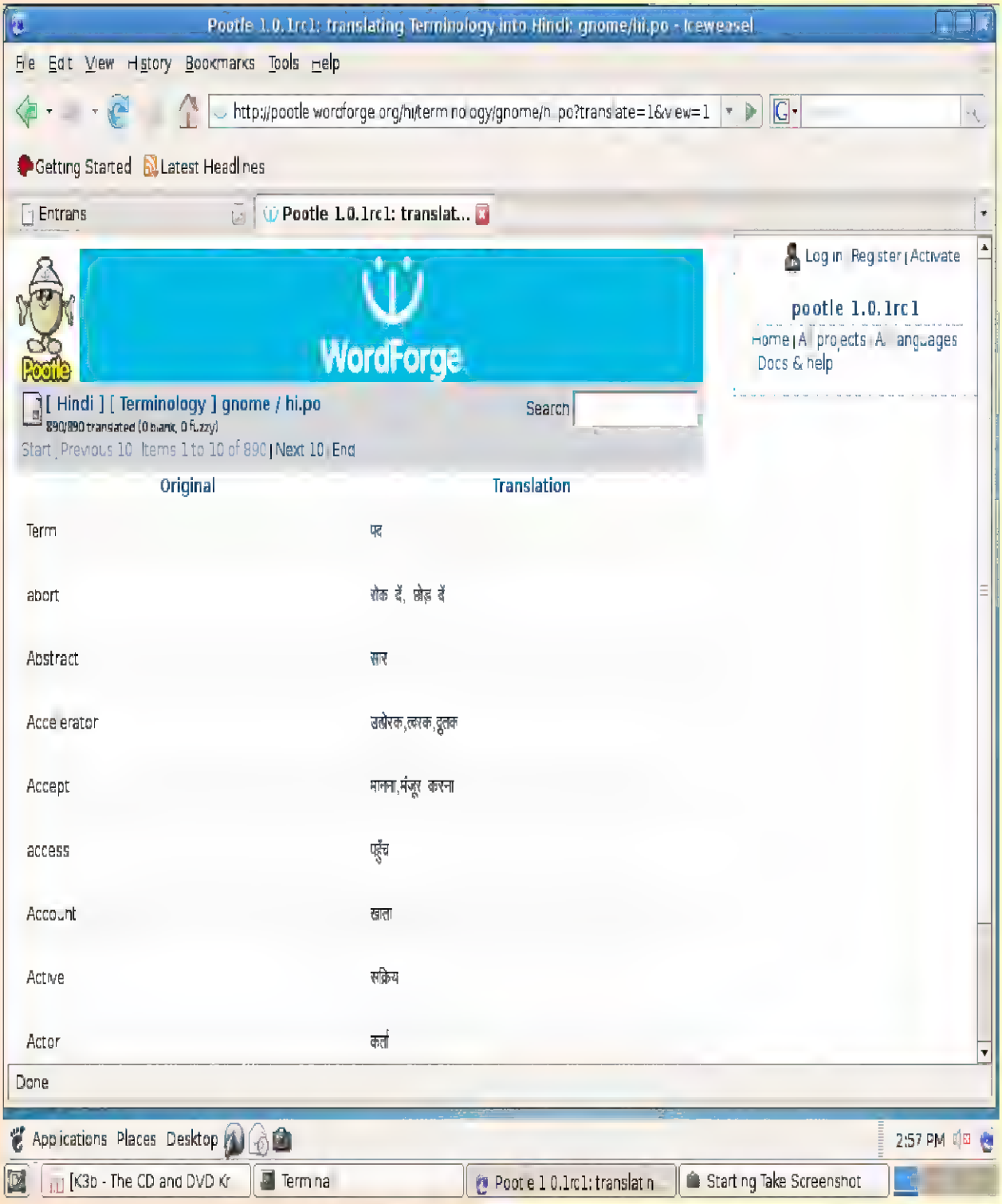
This tool is very useful if there are large numbers of interested contributors who can make small contributions regularly in their free time. It also avoids any need of complex setup; all that is needed is a browser and input methods to type in their language, both of which are now available on all platforms.



A live installation is available at http://www.indlinux.org/entrans/

C. Pootle

Pootle is another online translation management tool, allowing multiple teams to work on a common pool of strings drawn from different open source projects. While it does not give rough translation feature, it gives a dictionary lookup of translated terms. Apart from that it has lot of facility and tools to integrate with software projects, details of which are beyond scope. This tool is very easy to setup and can be used standalone or in a networked environment. It is best used for conducting translation marathons [2] where a large number of people can come and translate large amount of work in a short period of time.



A live setup is available here - http://pootle.wordforge.org/

V. CASE STUDY - INDLINUX EXPERIENCE OF TRANSLATIONS

Taking Indlinux Hindi as an example, here is the quick recount of Linux localization:

Somewhere around 2002, Gnome had shown capabilities for initial support for Indic Unicode characters – typically, Hindi. This led to the dream of having full fledged Linux OS in Indic languages. Things were got assembled from ground zero – and bits and pieces were put together to work. From fonts to rendering engines,

bugs were everywhere. Work on steady pace was going on for about at least two years without any visible progress in hand. Indlinux mostly concentrated in Hindi, and later on, other team started on their own - taking inputs and help from Indlinux. Ankur Bangla, PunLinux Panjabi, Utkarsh Gujarati etc were such teams. Since translations were done by volunteers from all across India, translated terms varied as per local usage. This led to the need of translation workshops that helps weed out translation errors, out of context translations, inconsistencies, etc. Hindi translation got a shot in the arm when Sarai sponsored initial project to translate GNOME 2.0 in Hindi, and later, KDE 3.2 and OpenOffice 2.0 Help. Sarai was also a pioneer in initiating and pushing Hindi translation workshops to refine translations and to train translators. A Hindi glossary was also evolved to form a standard for common technical terms used. The workshops were of immense help in giving ideas- what to translate and what not to, and coining new words. What we have today, a refined Hindi Linux Desktop might not have been possible without the help of Sarai's sponsorships. From this experience to get results two things were clear

1. Sponsor translator and translation team at least those who can deliver,
2. Organize workshops to translate/refine existing translation en-mass; without these, the achieved progress may not be up to the mark.

In the initial days of localization, translators did not have much option for their translation tool. They were bound to use plain odd Yudit the Unicode editor. Every string needed to be translated and typed at every occurrence however, umpteen times they were repeated. That made translation works very slow, boring, repetitive, and taste-less. Things changed dramatically when translation tools like Gtranslator, Kbabel, Poedit came in picture. As the work grew, with the help of translation database and facility of machine added auto translations, productivity as well as consistency in translation was maintained at high level. Soon, under Indlinux a multilingual live Linux CD Rangoli was released that clearly showed the potential of local language computing.

At a later stage, online translation tools like Rosetta, Entrans also made available. However, they failed to attract Indic translators much, may be due to their tedious and time consuming, slow responsive interfaces. Anyway, online translation tools like Entrans were the need of the day. They will ultimately find their way and most translations will be done through these tools in near future.

VI. A COMPARISON OF TERMINOLOGY EFFORTS

A small compilation of translated terminologies evolved by three independent efforts is given below. Due to space limitations only a small set is taken, which may not prove well benefits of open source process, since language being the same, translations can ultimately emerge to be similar or same through constant process of review. The comparison is between user interface terminology from Hindi version of Windows XP, and IT terms glossary provided by CSTT-CDAC and the translations reached by IndLinux Hindi translation effort.

Notes: Blank entries indicate either there was no translation listed or could not be searched upon easily or an occurrence not found in the interface reviewed.

	XP Hindi	CSTT	IndLinux
Accessories	सहायक उपकरण	सहयंत्र	सहायक उपकरण
Active	सक्रिय	सक्रिय	सक्रिय
Add	जोड़ें	योजी स्कलन योग	जोड़ें
Address bar	पता पट्टी		पता पट्टी

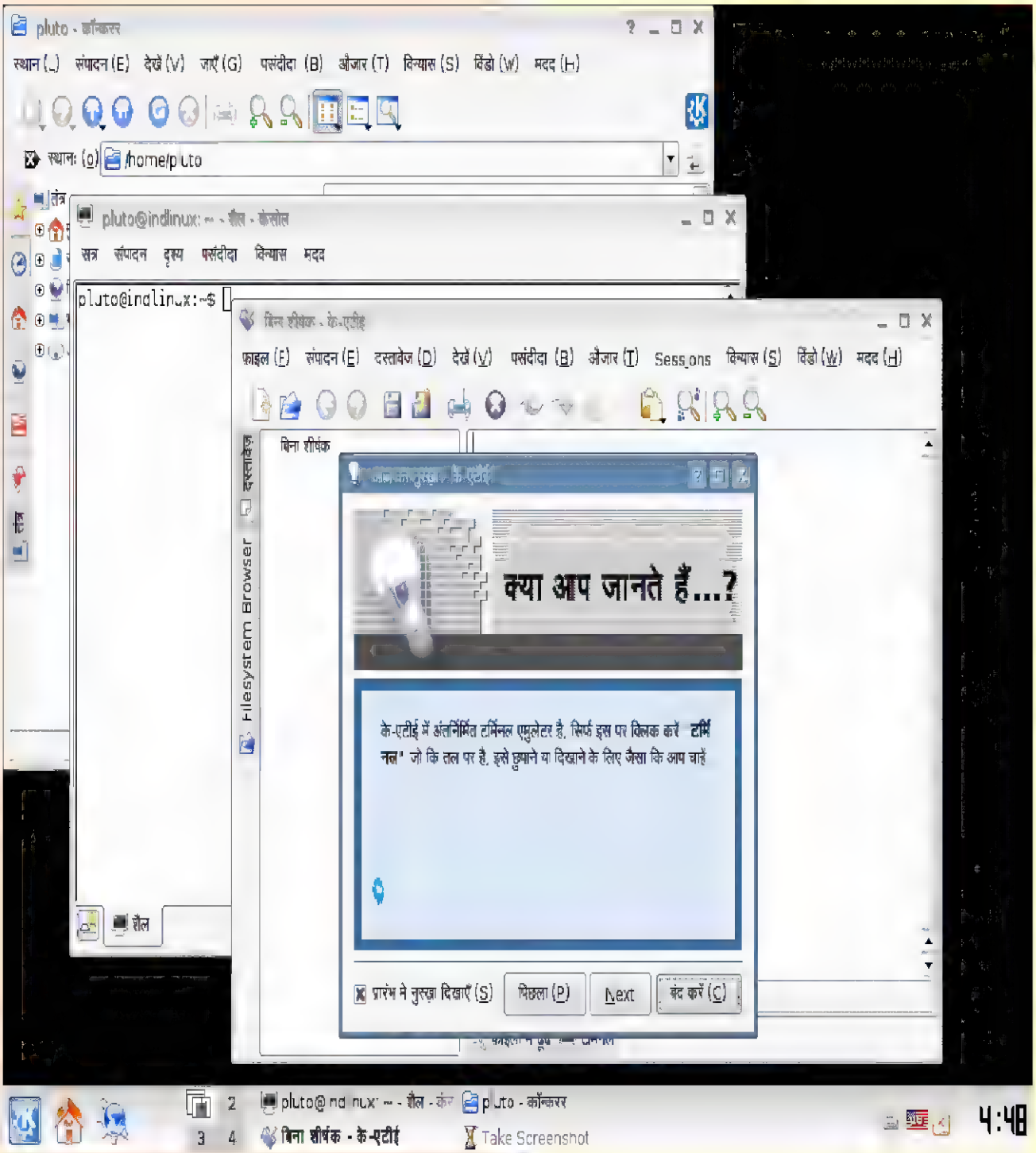
Address book	पता पुस्तिका	पता पुस्तक	पता पुस्तिका
Align	संरेखित	सरेखण	कतारबद्ध,पंक्तिबद्ध करें,सरेखण
Appearance	प्रकटन		रूप ,प्रकटन
Apply	लागू करें		लागू करें
Apply	लागू		प्रयोग करें

Arrange	व्यवस्थित		जमाएँ, व्यवस्था
Automatic	स्वचलित	स्वचालित	स्वचलित
Back	वापस	पश्च	वापस, पिछला
Background	पृष्ठ भूमि	पृष्ठ भूमि	पृष्ठभूमि
Bar	पटटी	स्तम्भ, रेखिका	पट्टी
Bookmark	पसद	पृष्ठित स्मृति	पुस्तचिह्न, पसद
Cancel	रद्द करें	निरसन	रद्द
Centred	केंद्रित	-	
Change	बदलें		बदलें
Clear	साफ करें	रिक्त	साफकरें
Close	बद	बन्द	बंद करें
Comments	टिप्पणियाँ	टिप्पणी	
Control panel	नियत्रण कक्ष	नियत्रण पट्टिका	
Copy	प्रतिलिपि	प्रतिलिपि	प्रतिलिपि, नकल
Customize	अनुकूलित करें		फरमाइशें
Default	डिफाल्ट	डिफॉल्ट	डिफॉल्ट, नयशुदा, सुनिश्चित
Details	विवरण	विस्तार, विवरण	
Display	प्रकटन	प्रदर्श	प्रदर्शक प्रदर्शन
Document	दस्तावेज	प्रलेख	दस्तावेज
Edit	सपादन	सपादन	सम्पादन
Empty	रिक्त	रिक्त	खाली, रिक्त
Existing	मौजूदा	विद्यमान	मौजूदा
Exit	बाहर निकलें	निर्गम	बाहर, निर्गम, निकास
Expand	विस्तारित	विस्तारित	फैलाएँ, विस्तार
Favourites	पसद	पसदीदा	पसदीदा
Find	ढूढें	अन्वेषण	ढूढें
File		सचिका	फाइल
Format bar	स्वरूप पट्टी	सरूप	
Forward	अग्रेषित करें	अग्र	आगे बढाएँ, अग्रेषित

Go to	इस पर जाए		यहां पर जाएँ, जाएँ
Help	मदद	सहायता	मदद
Home	मुख	आमुख, निजी	अथ, इब्तदा, आशियाना, घर, शुरुआत
Home page	मुख पृष्ठ	आमुख पृष्ठ, मूल स्थान पृष्ठ	मुख पृष्ठ
Icons	चिन्ह		प्रतीक, चिह्न
Invert selection	चयन पलटें		चयन पलटें, चुने हुए को उलटें
Keyboard	कुंजीपटल	कुंजीपटल, कीबोर्ड	कीबोर्ड,कुजीपट
List	सूची	सूची	सूची
Lock	अवरोधित	पाश, लॉक	ताला
Notification	सूचनाएं		अधिसूचना
Ok	ठीक		ठीक
Open	खोलें	खुला	खोलें,खोलना,खुला,मुक्त
Option	विकल्प	विकल्प	विकल्प
Page setup	पृष्ठ सेटअप	पृष्ठ	पृष्ठ सेटअप, पृष्ठ विन्यास
Paragraph	अनुच्छेद	पैराग्राफ	अनुच्छेद, पैराग्राफ
Paste special			विशेष चिपकायें
Print	मुद्रण		छापें,छपाई,मुद्रण
Print preview	मुद्रण	मुद्रण	छपाई नमूना,

	पूर्वावलोकन		छपाई पूर्वावलोकन
Property	गुण		गुण
Re do	दोहराएं		दोहराएँ, पुन: करें, फिर से करें
Refresh	ताजा करें	पुनर्चर्या	ताजा करें
Registered	पंजीकृत	पजीकृत	पजीकृत
Remove	हटाए		मिटाएँ
Rename	नाम बदलें	पुन:नामकरण	पुनर्नामकरण,नाम बदलें,नया नाम
Replace	बदलें		बदलें
Restart	पुनरारभ		फिर प्रारंभ करें
Restore	पुनर्स्थापित करें	पुन:स्थापन	बहाल करें, पुरानी स्तिथी में लाऐ, यथावत करें
Ruler	मापनी		
Save	सहेजें	सुरक्षित करो, बचाओ, सेव	सहेजें, संग्रहितकरें,सचित करें
Save as	इस रूप में सहेजें		ऐसे सहेजें
Search	खोजें	खोज	ढूढें , खोजें
Security	सुरक्षा	सुरक्षा	सुरक्षा
Select all	सभी का चयन करें		सभी चुनें, सबको चुनें
Send to	भेजें		

Share	साझा	शेयर	साझा, साझेदारी
Shared	साझा		साझा, साझेदारी
Standard	मानक	मानक	प्रमाणिक,मानक, प्रामाणिक
Start	प्रारभ	प्रारभ	प्रारभ,शुरू,चालू
Start menu	प्रारभ मेनू		स्टार्ट मेन्यू
Status bar	स्थिति पट्टी	अवस्थिति	स्थिति पट्टी
Stop	रूकें	विराम	रोकें, रूकें, बन्द
Taskbar	कार्य पट्टी		कार्यपट्टी
Text wrap	पाठ लपेटें		
Tip	युक्ति		युक्ति, सकेत
Tools	उपकरण	उपकरण	उपकरण
Un do	पूर्ववत करें		पहले जैसा, अकृत करें,
Untitled	अनामांकित		बेनाम,अनाम
Up	ऊपर		ऊपर
Up one level	एक स्तर ऊपर		एक स्तर ऊपर
Urgent	अत्यावश्यक		अत्यावश्यक, तत्काल
User	उपयोगकर्ता	उपयोगकर्ता, प्रयोक्ता	उपयोक्ता,प्रयोक्ता, कर्ता
View	दृष्य	दृश्य	देखें, दिखाएँ, दर्शन, नजारा
Welcome	सुस्वागतम		सुस्वागतम, स्वागतम्
Window	विंडो	विंडो	विंडो
Zip	संपीडित	सकुचित	
Zoom	ज़ूम	ज़ूम	ज़ूम,छोटा-बडा करें



VII. CONCLUSION

While what has been discussed may not be conclusive enough to prove that open source model of translation is better than a closed one, but the model and tools in offer give flexibility for innovation and improving the process of building translations, lexicons, etc. A case in example is Wiki and the content management systems which allow anyone to post , review content, with moderation strings attached. Wikipedia has quickly become the largest online encyclopedia. Also Wiktionary.org, which bases itself on wiki, is evolving a multilingual lexicon and dictionary.

The open source tools like entrans, pootle, etc can be customized to build dictionaries or lexicons, which can be developed in a collaborative way,

drawing in a large pool of contributors. While a closed door model may bring in good quality by having experts do the work, open model benefits from the volumes and that it is the users of the final product who have a say from the beginning, which means the lexicon has a wider acceptance and is more in tune with the prevalent language usage and vocabulary. While it may not satisfy the purist's view, but ultimately, if a language is not open to accept & accommodate, it may well be on its path to oblivion.

ACKNOWLEDGMENT

Many thanks to Ravikant, Gora Mohanty, Alka Irani, and Shashi Palekar for the useful feedback and suggestions in writing this paper. Also thanks to CDAC Mumbai and CSTT for organizing the seminar.

REFERENCES

[1] http://www.indlinux.org
[2] Online Hindi translation - Entrans - http://www.indlinux.org entrans
[3] Pootle - http://pootle.wordforge.org
[4] Wiktionary - http://www.wiktionary.org
[5] GNU gettext -http://www.gnu.org/software/gettext

This paper was presented at LRIL-2007- National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.5 Statistical Analyses of Myanmar Corpora

Hla Hla Htay, G. Bharadwaja Kumar, and Kavi Narayana Murthy

Abstract Myanmar (Burmese) is a member of the Burmese-Lolo group of the Sino-Tibetan language spoken by about 21 million people in Myanmar (Burma). Corpus based statistical analyses have not yet taken much focus on Myanmar language. In this paper, we have carried out statistical analysis on Myanmar text corpora of about 1.6M words [150,000 sentences]. In Myanmar language, sentences are clearly delimited by a unique sentence boundary marker but words are not always delimited by spaces It is therefore non-trivial to segment sentences into words. Syllabification itself is a non-trivial task. Syllabification is done with longest string matching using syllable list. Word Segmentation is performed with longest syllable matching by looking up the dictionary. We have checked manually the outputs of syllabification and word segmentation modules. We have obtained 99% accuracy on syllabification and 80% on word segmentation. We have carried out growth rate analysis and coverage analysis of Myanmar corpus. We have also carried out analyses of word length and sentence length distributions of Myanmar corpus.

Index Terms—Corpora, Coverage, Entropy, Myanmar, Syllabification, Tokenization, Type-Token,

I. INTRODUCTION

MYANMAR (Burmese) is a member of the Burmese-Lolo group of the Sino-Tibetan language spoken by about 21 million people in Myanmar (Burma). According to history, Myanmar script has originated from Brahmi script which flourished in India from about 500 B.C. to over 300 A.D. The script is syllabic in nature, and written from left to right.

Myanmar script is a writing system constructed from consonants, consonant combination symbols (i e. Medials), vowel symbols related to relevant consonants and diacritic marks indicating tone level (niggahita, visajjaniya). Myanmar language is composed of 33 consonants, 11 consonant combination symbols, 12 basic vowels and extension vowels, vowels symbols, diacritic marks and devowelizing consonants [1], [2]. See in Table 1 and 2. Sentences are clearly delimited by a unique sentence boundary marker "။" which is called ပုဒ်ဖ [pou ma.] but words are not always delimited by spaces. Although there is a general tendency to insert spaces between phrases, inserting spaces is more of a convenience rather than a rule. Spaces may sometimes be inserted between words and even between a root word and the associated postposition. In fact in the past spaces were rarely used. It is therefore non-trivial to segment sentences into words. Word tokenizing plays a vital role in most Natural Language Processing applications such as Summarization, Information Retrieval, Machine Translation, Text Categorization, and so on

Table-I
CONSONANTS

Consonants				
က [ka.] k	ခ [kha.] c	ဂ [ga.] *	ဃ [ga.] C	င [nga.] ɪ
စ [sa.] n	ဆ [hsa.] e	ဇ [za.] Z	ဈ [za.] ns	ည [nja.] n
ဋ [ta.] #	ဌ [hta.] q	ဍ [da.] !	ဎ [la.] ɹ	ဏ [na.] P
တ [ta.] w	ထ [hta.] x	ဒ [da.] '	ဓ [da.] "	န [na.] e
ပ [pa.] y	ဖ [hpa.] z	ဗ [ba.] A	ဘ [ba.] b	မ [ma.] ;
	ယ [ja. or .a.] &	ရ [ja. or .a.] [	လ [la.] V	ဝ [wa.] 0
	သ [tha.] o	ဟ [ha.] [	ဠ [la.] V	အ [a.] t
Consonants Combination symbols				
္ -s ျ -G ြ -S ွ -R ှ >ု -Q ိ -. ီ -, ံ -I ့ -W				
Devowelizing consonants				
-် -ါ -ာ -ိ -ီ -ု -ူ -ေ -ဲ -ဳ -ဴ -ဵ -ံ -့ -း -္ -် -ါ -ာ -ိ -ီ -ု -ူ -ေ -ဲ -ဳ -ဴ -ဵ -ံ -့ -း -္				
် Xf က Pf ာ wf ိ xf ီ df ု "f ူ cf ေ uf ဲ Af				
ဳ bf ဴ rf ဵ ,f ံ &f ့ vf း 0f ္ of ် [f ါ Vf				

Table-II
VOWELS AND ITS EXTENSION

ာ[a]ါ[a]tm?ိ [ı] ီ[ı]£ı?[u]Oı?[u]ODı ေ[eı]{?ဲ[eı]tJ?ံ[o]Mo?ံ[o]aMomf?့[an]tH?း[ou]tdk		
- -	် -m	်း -m,
့ -d	့ -D	့း -D;
် -k	် -l	်း -l,
ော a-	ော် a-h	ော်း a-,
ာ -J	ာ် -Jh	ာ်း -Jh,
ောာ a-m	ောာ် a-mh	ောာ်း a-mf
ား -H	ား် -Hh	ား်း -Hh,

Corpus based approaches to language have made significant contributions to linguistic research as also in education and language technology. A corpus is a large and representative collection of language material stored in a computer processable form [3]. Corpora provide realistic, interesting and insightful examples of language use for theory building and for verifying hypotheses [4], [5], [6], [7]. Insights obtained from analysis of corpora have led to fresh and better understanding of how language actually works [8], [9], [10], [11]. Corpora provide the basic language data from which lexical resources such as dictionaries, thesauri, word-nets, etc. can be generated. Language technologies and applications such as Morphological Analyzers, Stemmers, Syntactic Parsers, Spell Checkers, Information Retrieval systems, Information Extraction systems, Automatic Text Summarization systems, Automatic Text Categorization systems,

Hla Hla Htay, G. Bharadwaja Kumar and Kavi Narayana Murthy are with the Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andera Pradesh, 500046 (e-mail: hla hla htay@yahoo.co.uk, g vijayabharadwaj@yahoo.com, knmuh@yahoo.com)

Machine Translation systems etc. greatly benefit from language corpora. Development of large and representative corpora and annotating them with morphological, syntactic and semantic information is therefore considered to be a priority area. Corpus based statistical approaches have emerged as promising alternatives to traditional linguistic approaches. Hybrid approaches that combine traditional linguistic approaches with corpus based statistical approaches have also become attractive

Corpus based studies in English date back to 1960s [12], [13], [14], [15], [16], [17]. Corpus linguistics has not yet become a major aspect of education and research in linguistics in Asian languages [18] in general and Myanmar in particular. Even plain text corpora available are inadequate and annotated corpora are hardly available in many languages. Even, a large scale plain text corpus is not available in Myanmar language. There is a greater need to develop large scale Myanmar corpora. In this paper, we will discuss briefly about the development of corpus carried out in University of Hyderabad. Then, we will explain algorithms for Syllabification and word tokenization. Finally, we will discuss in detail about the different analyses carried out on Myanmar corpus.

II. MYANMAR TEXT COLLECTION

Development of lexical resources is a very tedious and time consuming task and developing purely with manual effort is very slow. We have downloaded Myanmar texts from various web sites including news sites and on-line magazines. As of now, our Myanmar corpus contains 17 lakh sentences. The downloaded corpora need to be cleaned up to remove hypertext markup, etc. We have developed the necessary scripts in Perl. Also, different sites use different font formats and character encoding standards are not yet widely followed. We have mapped these various formats into the standard Winninnwa font format We have stored the cleaned up texts in ASCII format. This will enable processing in environments where Unicode is not yet supported It is easy to switch to Unicode where required

The corpus includes over 300 full books as well as free and trail text from online book store to include a wide variety of Myanmar writings including a variety of genres, types and styles - modern and ancient, prose and poetry. It also contains text from official newspapers in Myanmar. Text is converted to standard Winninnwa font using tools developed here. The corpora are ISCII encoded and are seen to be reasonably clean. A corpus should be constructed in keeping with the principles of corpus linguistics [19], [20], and [21]. It must be 'large' and 'representative'. A balanced corpus, however, does not mean nearly equal amounts of material from various genres, types and styles. Application of language in areas other than literature is a relatively recent phenomenon. Newspapers cover a wider variety of topics and styles including sports, science and technology, politics, economics and business, cinema, etc. No corpus should be put to use for a given application without a careful analysis of its nature and contents. While there can be no guarantee that our corpus is good enough for any given use or application, we feel that the corpus is good enough for some kinds of applications we have in mind. Our aim shall be to strive to build larger, more balanced and more representative corpora

Preliminary studies suggested that Myanmar sentences can be tokenized by eliminating stop words Hopple [22] also notices that particles ending phrases can be removed to recognize words in a sentence. Stop words are defined as non-information-bearing words They form closed classes and hence can be listed. Stop words include prepositions post-positions, conjunctions, particles, inflections etc. These words appear so frequently that their usefulness is limited. In Information Retrieval, for example, search engines ignore stop words at the time of searching a key phrase In Information Extraction and Text Summarization also, stop words are pushed aside and treated as irrelevant information, in order to extract the most relevant and important information We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of English-English- Myanmar (Students Dictionary) [23], English- Myanmar Dictionary [24], and The Khit Thit English-Myanmar dictionary [25] We have also looked at stop word lists in English [26] and mapped them to equivalent stop words in Myanmar See Table 3. As of now, our stop words list contains about 1216 entries Stop words can be prefixes of other stop words leading to ambiguities. Usually, the longest matching stop word is the right choice. We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of English-English-Myanmar (Students Dictionary) [23], English-Myanmar Dictionary [24], and The Khit Thit English-Myanmar dictionary [25]. We have also looked at stop word lists in English [26] and mapped them to equivalent stop words in Myanmar. See Table 3. As of now, our stop words list contains about 1216 entries. Stop words can be prefixes of other stop words leading to ambiguities. Usually, the longest matching stop word is the right choice.

As we keep analyzing texts, we can identify some words that can appear independently without combining with other words or suffixes. We build a list of such valid words and we keep adding new valid words as we progress through our segmentation process, gradually developing larger and larger lists of valid words This list of known words can be made use of for hypothesizing candidate words as we go along

Myanmar language uses a syllabic writing system [27] unlike English and many other western languages which use an alphabetic writing system. Interestingly, almost every syllable has a meaning in Myanmar language. This can also be seen from the work of Hopple [22]

TABLE-III
STOP WORDS OF ENGLISH VS MYANMAR

Prepositions and adverbs	
always	အမြဲ [a mje], အမြဲတမ်း [a mje dan], အမြဲတစေ [a mje da zei]
Nominative personal pronouns	
I	ကျွန်တော် [kjun do], ကျွန်မ [kja ma], ငါ [nga], ကျွန် [kjou], ကျနော် [kja no], ကျွန်ုပ် [kja nou'], ကျမ [kja ma]
Accusative personal pronouns	
me	ကျွန်တော်အား [kjun do a:], ကျွန်တော်ကို [kjun do kou], ကျွန်မကို [kja ma gou] , ငါ့ကို [nga gou], ကျွန်ုပ်ကို [kjou kou], ကျွန်ုပ်ကို [kja nou gou]
Reflexive personal pronouns	
myself	မိမိကိုယ်တိုင် [mi mi kou dain], မိမိကကာ [mi mi hpa dha], မိငယ်ကို ယ်တိုင် [kou kou dain], ကို နှံကာသာ [kou hpa dha]
Relative pronouns	
That	သည့် [thi], မည့် [mji], တဲ့ [te]
Possessive pronouns and adjectives	
my	ကျွန်ုပ်၏[kja nou' i.], ကျွန်တော်၏[kjun do i.], ကျွန်မ၏[kja ma i.], ကျနော်၏[kja nou i.], ကျမ၏[kja ma i.], ငါ့ရဲ့ [nga ɪ.], ကျွန်ုပ်ရဲ့ [kjou i.], ကျွန်ုပ်ရဲ့ [kja nou' je.], ကျွန်တော့်ရဲ့ [kjun do je.], ကျွန်မရဲ့ [kja ma je], ကျနော့်ရဲ့ [kja no je], ကျမရဲ့ [kja ma je], ငါ့ရဲ့ [nga je], ကျွန်ုပ်ရဲ့ [kjou je.], ကျွန်တော့် [kjun do], ကျနော့် [kja no.]
Demonstrative pronouns and adjectives	
this	ဒီဇာ [i a ja], ဟောဒါ [ho da], ဟောဒီ [ho dhi], ဒီ [dhi]
Indefinite pronouns and adjectives	
some	အချို့ [a chou], အချို့သော [a chou tho], တချို့ [ta chou.], တချို့သော [a chou. tho], တချို့ချို့ [ta chou.ta chou.], တချို့တလေ [ta chou ta lei]
Conjunctions	
and	နှင့် [hnin], ပြီးလျှင် [pji hljin], ၎င်းနောက် [la gaun]
Questions	
why	အကဲ ဘိတဲ့သို့ [a be khe dho.], မည်ကဲ့သို့ [mji khe dhou], မည်ၾ ည့်နည်းနှင့် [mji dhi ni hnin.] , မည်ၾ ည့်နည်းဖြင့် [mji dhi ni. hpji 0], မည်သို့ [mji dhou], ကဲလိုလဲ [be lou le'], သို့ မဟုတ် [dho bei me.], မည်ၾ ည့်နည်းနှင့်မဆို [mji dhi ni hnin ma hsou], ကဲ့နည်းနှင့် [be ni' hnin], မည်ရွေ့မည်မျှ [mji jwei mji hmja], အကဲ ဘိမျှလောက် [a be hmja lau 0], ကဲလောက် [be lau 0]

We have developed scripts in Perl to syllabify words using our list of syllables and then generate n-gram statistics using Text::Ngrams, which is developed by Vlado Keselj [28] Examples of collected Ngrams are shown in Table 4. We have used "type word" option treating syllables as words. We had to modify this program a bit since Myanmar uses zero (as ဝ [wa] letter) and the other special characters ("-", "<", ">", ".", "&", "[", "]" etc) which were being ignored in the original Text::Ngrams software We collect all possible n-grams of syllables up to 5-grams. Almost all monograms are meaningful words. Many bigrams are also valid words and as we move towards longer n-grams, we generally get less and less number of valid words. We have used mutual information for even-syllables words and maximum entropy for odd-syllables to hypothesize possible words. Manual checking is essential to finally choose valid words

There are lots of valid words which are not described in published dictionaries. The entries of words in the Myanmar-English dictionary which is produced by the Department of the Myanmar Language Commission are mainly words of the common Myanmar vocabulary. Most of the compound words have been omitted in the dictionary [1]. This can be seen in the preface and guide to the dictionary of the Myanmar-English dictionary produced by Department of the Myanmar Language Commission, Ministry of Education. 4-syllable words like "ထူးထူးဆန်းဆန်း" [htu: htu: zan zan:] (strange), "ထူးထူးကဲကဲ" [htu: htu: ke: ke:] (outstanding) and "ထူးထူးခြားခြား" [htu: htu: gja: gja:] (different) are not listed in dictionary although we usually use those words in every day life. Statistical construction of machine readable dictionaries has many advantages. New words which appear from time to time such as internet, names of medicines, can also be detected Compound words also can be seen.

TABLE-IV
EXAMPLES OF COLLECTED NGRAMS

No.	Bigram bisyllables	Tngram 3-syllables	4-gram 4-syllables
1	ဖန်ထည် glassware [hpan de]	ဝါးခန laughing or yawning loudly [wa ga ne:]	ငယ်ငယ်ကြီးကြီး young or old [nge nge kji kji]
2.	ဖန်ကျိုး glass stone [hpan toun:]	ဂုန်းခန် with an uproar [woun ga ne]	ခါးခါးသီးသီး bitterly [kha ga. thi dhi]
3.	ဖန်တီး create [npan di:]	ဃ၀်ခန with an roar [wo ga ne:]	ကောင်းကောင်တင်းတင်း very stout [taun. taun din din.]
4.	ဖန်လာ happen [hpan la]	အားခန a loud voice [a. ga ne:]	နှစ်နှစ်ကာကာ like much [hni' hni' the' the']
5	ဖန်အိမ် lantern [hpan e.n]	ဂုန်းခန with a big sound [boun. ga ne:]	နှစ်နှစ်ကာထား whole-heartedly [hni hni ka ga]
6.	ဖန်ဆင်း create [hpan zin']	ဇွေခန with supernat-ural power, with squinted eye [swei ga ne]	ဖျစ်ဖျစ်နှစ်နှစ် thick [pji pji hni hni]
7.	ဖန်သား glassware [hpan tha]	ဇွေ့ခန effortlessly [swei ga ne]	ပို့ပို့ဝါးဝါး outbravely [wun wun sa za]
8.	ဖန်ခုန် game's name [hpan khoun]	ဆောင့်ခန with stamping [hsaun. ga ne:]	စွန့်စွန့်စားစား riskly [sun. sun. sa: za:]
9	ဖန်ခွက် glass [hpan gwe 0]	မှေးခန a short nap [hmei ga ne:]	စဉ်းစဉ်းစားစား thoughtfully [sin' sin' sa za]
10.	ကန်ဧရာ bank of lake [kan saun:]	ထောင်းခန fuming with rage [htaun ga ne:]	များများစားစား many,much [mja mja sa: za]

With this technique, morphological structure of words can also be analyzed. See in Table 5. The above-mentioned three and four-syllable words are adverbs derived from the verbs "ထူးဆန်း" [htu: zan:], "ထူးကဲ" [htu: ke:], and "ထူးခြား" [htu: gja:]. Statistical dictionaries can be updated much more easily than published printed dictionaries, which need more time, cost and man power to bring out a fresh edition. Common names such as names of persons, cities, committees etc. can also be mined. Length statistics will be a useful hint and many researchers have used longest string matching [29], [30]

TABLE-V
EXAMPLES PATTERNS OF MYANMAR MORPHOLIGICAL ANALSYSIS

A	B	C	D	F
basic unit syllable	(Verb) ကောင်းသည်	(Noun) အကောင်း	(Negative) မကောင်းကျး	(Noun) ကောင်းမှု
ကောင်း [kaun] good (Adj)	[kaun dhi] is good	[a kaun] good	[ma kaun. bu:] Not good	[kaun mhu.] good deeds
ဆိုး [hso:] bad (Adj)	[hso: dhi] is bad	[a hso:] bad	[ma hso: bu:] Not bad	[hso: mhu.] Bad Deeds
ရောင်း [jaun:] sell (Verb)	[jaun: dhi] sell	[a jaun:] sale	[ma jaun. bu.] not sell	[jaun. mhu.] sale
ရေး [jei.] write (Verb)	[jei. dhi] write	[a jei.] writing	[ma. jei bu:] do not write	[jei mhu.]
ပြော [pjo.] talk,speak (Verb)	[pjo. dhi] talk,speak	[a pjo.] talk,speech	[ma pjo bu:] not talk,speak	[pjo mhu.] a talk

III. SYLLABIFICATION AND WORD SEGMENTATION

Since dictionaries and other lexical resources are not yet widely available in electronic form for Myanmar language, we have collected possible syllables (including ဝမ္မ , တက္ကသိုလ်) and 2 lakhs Myanmar word-lists. With the help of these stored syllables and word lists, we have done Syllabification and word segmentation. The first step to build a word hypothesizer is syllabification of the input text by looking up syllable lists. In second step, we exploit lists of words (n-grams at syllable level) for word segmentation from left to right

Myanmar Natural Language Processing Group has listed 1894 syllables that can appear in Myanmar texts. We have observed that there are some more syllables, especially in foreign words including Pali and Sanskrit words which are widely used in Myanmar. We have collected other possible syllables using Myanmar-English dictionary. As we collected texts from internet which has lack of standard typing sequences, we also collected different possible typing sequences of syllables which will be seen as same appearance. Following is an example of syllables in different typing sequences. Now we have over 4000 syllables in our list. Forming of syllables is growing longer and longer from left-to-right, e.g. u (က), um (ကာ), um: (ကား), aMum (ကြော), aMumf (ကြော်), aMumf (ကြော်), aMumifh (ကြောင့်), aMumif: (ကြောင်း), etc. We have developed scripts in Perl to syllabify words using Longest String Matching and our list of syllables

TABLE-VI
SYLLABLES WITH DIFFERENT TYPING SEQUENCES

On screen	တြီး ကြီး:	ကို ကို
In ascii	MuD, BuD,	udk ukd

In Myanmar Text Syllabification, Longest string matching alone can be handled. The table 6 also shows that different Typing sequences of syllables are detected Failure caused due to

1. the combination of the writing sequences (typing sequences) of syllables
2. words borrowed from foreign languages
3. the need of new syllables entries which are rarely used

A. *Longest String Matching*

In this matching, it goes from left-to-right scan in greedy manner.

```
1.  Load the set of syllables from syllable-file
2.  Load the sentences to be processed from sentence-file
3.  Store all syllables of length j in N where j = 10...1
4.      for each sentence do
5.          length → length of the sentence
6.          pos → 0
7.          while (length > 0) do
8.              for j = 10...1 do
9.                  for-each syllable in N do
10.                     if string-match sentence (pos, pos + j) with syllable
11.                         Syllable found. Mark syllable
12.                         pos→ pos + j
13.                         length → length - j
14.                     End if
15.                 End for
16              End for
17.         End while
18. Print syllabified string
19. End for
```

Similarly, we have done tokenization with longest syllable matching using collected 2 lakh words list.

```
1.  Load the set of words from word-file
2.  Load the sentences to be processed from sentence-file
3.  Store all words of length j in N where j = 10..1
4.      for each sentence do
5.          length → length of the sentence in terms of syllables
6.          pos → 0
7.          while (length > 0) do
8.              for j = 10..1 do
9.                  for each word in N do
10.                 if string-match sentence (pos, pos + j) with word
11.                     Word found. Mark word
12.                     pos → pos + j
13.                     length → length - j
14.                 End if
15.             End for
16.         End for
17.     End while
18. Print tokenized string
19. End for
```

Example sentence segmentation is given in Table 7. Even though *syllabification* can be done with longest string matching, *tokenization* has needed to be improved with Hidden Markov Models (HMM) like machine learning techniques in order to get perfect system. Research is still going on. We have achieved about 99% accuracy in *syllabification* and 80% accuracy in *word segmentation* [31].

TABLE-VII
A SENTENCE BEING SEGMENTED INTO WORDS

ကျောင်းအုပ်ဆရာကြီး ဟာ သူ့ကြမ်းတမ်မှုကို စက်ဆုပ် သည်				
ကျောင်းအုပ်ဆရာကြီး	ဟာ	အကြမ်းတမ်မှု	ကို	စက်ဆုပ် သည်
[kyaung. aop hsa ya kyi:]	[tha]	[a kyan: phak mhu]	[ko]	[sak sop thi]
The headmaster		violence		abhors
N_{subj}	Particle	N_{obj}	Particle	$V_{present}$

IV. PRELIMINARY ANALYSIS OF MYANMAR CORPUS

Here we present a preliminary analysis of Myanmar corpus developed in University of Hyderabad. This includes news, novels, online magazines, and free and trail text of online bookshop. We have approximately 17 lakh sentences (after reducing duplicates) but we have carried out analysis for only 1.6M words [150,000 sentences] in this work.

A. Type-Token analysis

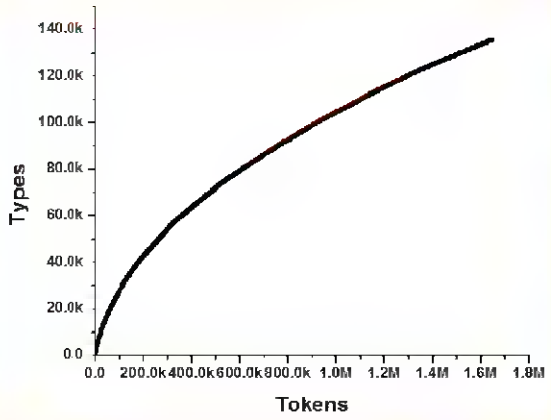


Fig. 1. Type-Token Growth Rate Analysis of Myanmar Corpora

The figure 1 shows the results of a type-token growth rate analysis. Each distinct word form is a type and each occurrence of a type counts as a token. If we analyze the entire corpus in one go, we will get the total number of types, total number of tokens and the global type-token ratio. Instead, if we perform type-token analysis incrementally, by starting with a small randomly selected part of the corpus and iteratively adding more texts randomly, we get a type-token growth rate curve that shows how many new types will be found as the corpus size increases.

Note that by types we mean fully inflected word forms, not root forms or citation forms found in dictionaries. Also, compounding will have their effect and the tokens we get do not necessarily correspond to the linguistic definition of a word understood in semantic terms. There is no automatic way to extract words based on meaning. Wide coverage, high performance, robust morphological analyzers are not yet available in most languages under study and here we restrict our analyses to full words. From figure 1, we can see that the curve is not saturated which tells us that we need to analyze more corpus to understand the behavior of Myanmar language.

B. *Coverage analysis*

TABLE-VIII
SELFCOVERAGE ANALYSIS OF MYANMAR CORPUS

%Coverage	Approx. No. of Types
50	1300
60	2700
70	5700
80	12200
85	18900
90	31300
95	60300
96	69600
97	86100
98	102000
99	119000
100	135518

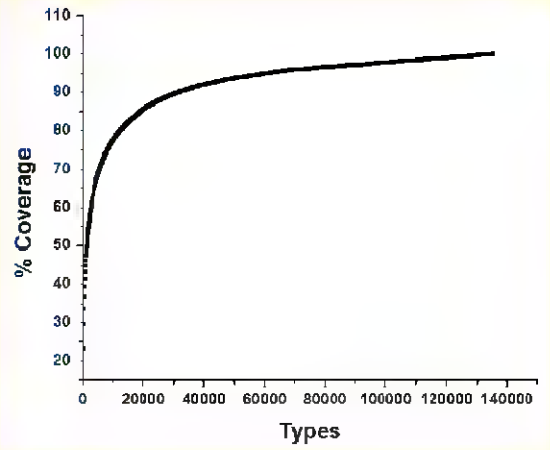


Fig. 2. Coverage Analysis of Myanmar Corpora

Coverage analysis deals with the examination of how much of a corpus can be covered by a given set of types. We perform a type-token analysis and prepare a list of types sorted in decreasing order of frequency of occurrence. By thresholding on this list, we can select the most frequent n words in the language, for any given value of n. We then explore what percentage of words in a corpus is found in the list so selected. Here we perform self-coverage analysis on the same corpus from which the words are extracted. (It would be instructive to perform coverage analysis on other corpora as and when they become available.)

From the figure 2, we can see that about 1300 most frequent words are sufficient to give about 50% coverage of the corpus. 60% coverage can be obtained by just the first 2700 words or so. This being self-coverage analysis 100% coverage can be obtained by using all the words in the word list.

C. *Sentence analysis*

Figure 3 shows the sentence length distribution for Myanmar corpus.

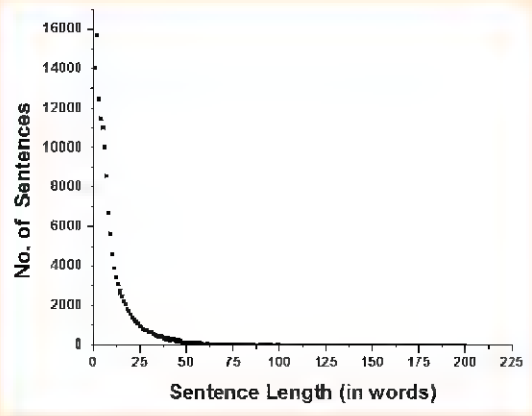


Fig. 3. Word Length Analysis of Myanmar Corpus

TABL- IX
WORDS AND SYLLABLE STRUCTURE

No of syllables	No. of words	Example
1	3776	ကောင်း
		Good (Adj)
		[kaun.]
2	45063	လိပ်ပြာ
		Butterfy, Soul (N)
		[.ei pja]
3	72795	ပြတင်းပေါက်
		Window (N)
		[ba din. bau]
4	48636	ပြည်တွင်းထုတ်ကုန်
		Domestic Product (N)
		[pji dwin. htou koun]
5	28932	လျှပ်စစ်ထမင်းအိုး
		Rice Cooker(N)
		[hlja si hta min. ou.]
6	16485	သူနာပြုဆရာမ
		Nurse(female) (N)
		[thu na bju. hsaja ma.]
7	9013	ရင်းနှီးမွန်းကြပေ တော့သ ည်
		become friend (V)
		[jin. hni. thwa. kya. pei to. thi]
8	4620	ပြည်ထောင်စုမြန်မာနိုင်ငံတော်
		Union of Myanmar (N)
		[pji daun zu. mja ma nain gan to]
9	2404	သ ဘာတ အရင်းအမြစ်များ
		Natural Resources (N)
		[than jan za ta. a jin. a mji]
10	1209	ခြေမလိုင်လက်မလိုင်ဖြစ်သ ည်
		be agitated or shaken(V)
		[chei ma_kain mi. le' ma_kain mi hpji thi]

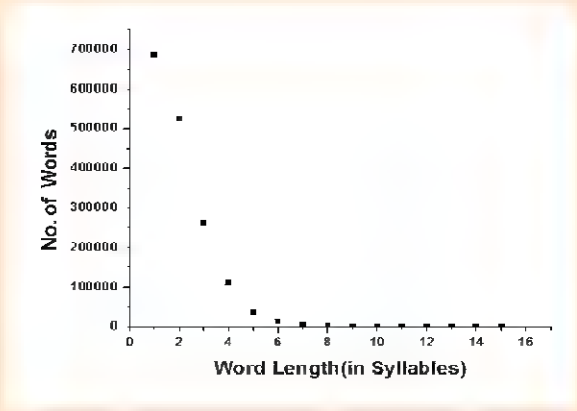


Fig. 4. Word Length in relation to Word Frequency

D. Word length variation with frequency

Figure 4 shows the word length distribution for Myanmar corpus. Words that occur frequently tend to be small words. It is therefore interesting to explore the relation between word frequency and word length. Figure 4 shows the scatter diagram of word length measured in word frequency. Word length is averaged over all words of a given frequency. It can be seen that the least frequent words are larger and word length shows a gradual decrease as we move towards more frequent words. High frequency words show a greater spread in terms of word length. Yet we can see trend - words tend to become smaller and smaller as we move towards the most frequent words. The streaks that we see are because of the clustering effect due to averaging

V. ENTROPY, PERPLEXITY

Entropy is a measure of information content. Entropy is related to probability, redundancy and uncertainty and is thus invaluable in language analysis. The more we know about something, the lower the entropy will be, because, we are less surprised by the outcome of a trial. Entropy can be interpreted as the minimum number of bits required to encode a given piece of information. Entropy can be calculated using the formula

$$H(X) = \sum_{x=1}^{N} p(x) \log_2 p(x)$$

where N is the number of Word Types in the language

H-maximum will be obtained when the probabilities of all the words in the corpus are same.

$$H_{\max} = \log_2 N$$

$$H_{relative} = \frac{H_{actual}}{H_{\max}}$$

$$redundancy = \frac{H_{\max} - H_{actual}}{H_{\max}}$$

Perplexity is useful for evaluating language models. A perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step. Perplexity of the given

model can be evaluated by

$$P(x) = 2^{H(x)}$$

where H(x) is the entropy of the given model.

The values of the Entropy and Perplexity for the corpus are shown in table 10.

TABLE-X
ENTROPY ANALYSIS OF MYANMAR CORPUS

Entropy	13.247
Relative Entropy	0.777
Redundancy	0.222
Perplexity	9725 432

VI. CONCLUSIONS

In this paper we have described syllabification which is important starting point to identify the words in the text, word tokenization, Myanmar stop-words, introducing Myanmar morphological formation and statistical analyses of a fairly large text corpus of Myanmar. Syllabification is done with longest string matching using syllable list. Word Segmentation is performed with longest syllable matching by looking up the dictionary. We have checked manually the outputs of syllabification and word segmentation modules. We have obtained 99% accuracy on syllabification and 80% on word segmentation. It is perhaps for the first time that a statistical analysis has been carried out on Myanmar language. These analyses point to issues relating to technology development as also to detailed linguistic analysis necessary for a complete understanding of the language. Larger corpora are needed in Myanmar language for meaningful analysis and technology development.

REFERENCES

[1] *Myanmar-English Dictionary*. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar

[2] Y. K. Thu and Y. Urano, "Text entry for Myanmar language sms: Proposal of 3 possible input methods, simulation and analysis," *in Fourth International Conference on Computer Applications*, Yangon, Myanmar, Feb 2006

[3] J. Sinclair, *Corpus, Concordance, Collocation* Oxford University Press, Oxford, 1991.

[4] M Barlow, "Corpora for theory and practice," *International journal of Corpus linguistics*, vol. 1, no. 1, pp. 1-38, 1996

[5] I. Lancashire, C. Percy, and C Mayer, *Synchronic Corpus linguistics*. Rodopi, Amsterdam, Atlanta, 1996

[6] W. Teubert, "Corpus linguistics: A partisan view," *International journal of Corpus linguistics*, vol. 4, no. 1, pp. 1-16, 2000

[7] N. Oostdijk and P Hann, *Corpus based research into language* Rodopi, Amsterdam, Atlanta, 1994

[8] D. Biber, "Investigating language use through corpus-based analyses of association patterns," *International journal of Corpus linguistics*, vol. 1, no. 2, pp. 171-198, 1996

[9] D. Biber, S.Conrad, and R. Reppen, *Corpus Linguistics: Investigating language structure and use*. Cambridge University Press, Cambridge, 1998

[10] C. Mair and M Hundt, *Corpus Linguistics and Linguistics theory*. Rodopi, Amsterdam, Atlanta, 2000

[11] M Stubbs, *Texts and Corpus analysis*. Oxford: Blackwell publishers, 1996

[12] J. Aarts and W. Meijs, *Corpus Linguistics: Recent development in the use of Computer corpora in English Language Research*. Rodopi, Amsterdam, Atlanta, 1984

[13] S. Johansson and A. B. Stenstrom, *English computer corpora: Selected papers and research guide* Mouton de Gruyter, Berlin, 1991.

[14] G. Knowles, B J. Williams, and L. Taylor, *A corpus of formal British English speech: The Lancaster IBM spoken English Corpus*. Longman, London, 1997.

[15] M. Ljung, *Corpus-based studies in English* Rodopi, Amsterdam, Atlanta, 1997.

[16] A. C F. Meyer, *English Corpus Linguistics*. Cambridge University Press, Cambridge, 2002

[17] R. Garside, G. Leech, and G. Sampson, *The computational analysis of English: A corpus based approach*. Longman, London, 1987

[18] G. B Kumar, K. N. Murthy, and B.B Chaudhuri, "Statistical analyses of telugu text corpora," To Appear in *International journal of Dravidian Langauges* (IJDL), vol. 36, no. 2, 2007.

[19] T. McEnery and A. Wilson, *Corpus Linguistics*. Edinburgh University Press: Edinburgh, 1996

[20] G. Kennedy, *An introduction to Corpus Linguistics*. Addison-Wesley, 1998

[21] D. Biber, "Representativeness in corpus design," *Literary and Linguistic computing*, vol 8, no. 4, pp 243-257, 1993

[22] P. Hopple, *The structure of nominalization in Burmese*, Ph.D thesis. May 2003

[23] *Student's english-english myanmar dictionary*, Ministry of Commerce and Myanmar Inforithm Ltd, Union of Myanmar, CD version, Version 1, 1999

[24] *English-Myanmar dictionary*, Ministry of Education, Union of Myanmar, CD version.

[25] S. U Soe, *The Khit Thit English-English-Myanmar Dictionary with Pronunciation*, Yangon, Myanmar, Apr 2000

[26] http www syger.com jsc docs stopwords english htm

[27] K. N. Murthy, *Natural Language Processing - an Information Access Perspective*, Ess Ess, ISBN- 81-7000-485-3, New Delhi, 2006

[28] V. Keselj, *Text ::ngrams*. http search cpan org vlado Text-Ngrams-1.8

[29] R. Anglell, G. Freund, and P Willett, "Automatic spelling correction using a trigram similarity measure," *Information Processing & Management*, vol. 19, no. 4, pp 305-316, 1983

[30] P. et al., "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants," *Information Research*, vol. 7, pp. 235-237, Jan 2001

[31] Hla Hla Htay, K. N. Murthy, "Myanmar word segmentation," in *Fourth International Conference on Computer Applications*, Yangon, Myanmar, pp. 353-357, Feb 2006

[32] Hla Hla Htay, K. N. Murthy, "Automatic Construction of Myanmar-English Bilingual Machine Readable Dictionary using Parallel Corpora", *Third International Conference On Computer Applications*, ,Yangon, Myanmar, page 667-671, Mar , 2005

[33] Hla Hla Htay, G. Bharadwaja Kumar, K. N. Murthy, "Building English-Myanmar Parallel Corpora", *Fourth International Conference on Computer Applications*, Yangon, Myanmar, page 231238, Feb, 2006

[34] B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished

This paper was presented at LRIL-2007, National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.6 Causative Compound Verb Constructions: A Generative Lexicon Account

Sanjukta Ghosh and Anil Thakur

Abstract Complex predicates are always in the focal area of research in both theoretical and computational linguistics due to their interesting syntactic and semantic behavior. One of the research questions in the literature of the complex predicate is where to account them: in the syntax or in the lexicon. Present paper provides a lexical treatment of one of the special type of complex predicates, viz., compound verbs used for causativization following Pustejovsky (1995).

Index Terms Causative Verbs, Complex Predicate, Generative Lexicon

I. INTRODUCTION

COMPLEX predicates are always in the focal area of research in both theoretical and computational linguistics due to their interesting syntactic and semantic behavior. One of the research questions in the literature of the complex predicate is how and where to account and accommodate them: in the syntax or in the lexicon. The present paper takes up one of the types of complex predicates, viz., the compound verbs used for causativization and presents their analysis within Generative Lexicon framework proposed by Pustejovsky (1995).

A compound verb can be defined as a verbal predicate composed of two verbs where the meaning of the compound is not compositional, i.e , it cannot be obtained by computing the meaning of their constituents. Because of this specific property of the compound verb and the complex predicate in general, they have always been considered as one of the major problems in many NLP applications. The complex predicates fall under the broader category known in Computational Linguistics literature as Multi Word Expressions (MWEs) which can be roughly defined as a sequence of words acting as a single unit at some level of linguistic analysis. Building lexicon for this kind of expressions is still not as developed as it is for single word expressions. However, the necessity for developing lexicons of different kinds of MWEs taking both syntactic and semantic information is recognized by the NLP research community worldwide especially in a multilingual environment. But as a matter of fact, this kind of project is still in the conceptual level in a truly multilingual society like India mostly because of the linguistic complexity associated with these expressions and lack of formal mechanism to handle that complexity in the mainstream linguistics. This paper is an attempt to formalize the lexical information of one particular kind of compound verbs for building lexical resources of Indian languages. The work can be extended for other kinds of compound verbs (where causativazation is not involved) and complex predicates ingeneral

II. BACKGROUND OF WORK

The languages of India (at least for instance, Indo-Aryan and Dravidian, the two major language families of India) have a large number of complex predicates. From theoretical linguistics point of view there have been attempts to analyze these constructions within Chomskyan generative grammar tradition as well as in some other theoretical models like Head-driven Phrase Structure Grammar (HPSG) and Lexical Functional Grammar (LFG). All these models of grammar attempt to generate the complex predicates by some syntactic operations in the grammar itself. Though LFG is a more lexicon-oriented framework, yet the operation adopted to change the predicate status is by introduction of a lexical rule in the grammar instead of a syntactic rule.

It is only in the last decade that a completely lexicon-centric model has been introduced by James Pustejovsky (1995). It differs from all the previous frameworks by virtue of its claim that the lexicon, like syntax, can also be generative. The Generative Lexicon framework is devised to account for the polysemous nature of words as well as their creative uses. It analyzes a word in four different structural levels: argument structure, event structure, qualia structure and lexical-inheritance structure. The argument structure in this framework not only provides the logical arguments of a predicate as discussed in other literature but also other possible arguments such as shadow arguments etc. Event structure is the definition of the event type (like state, process, transition) of a lexical item or a phrase. The framework also provides mechanism to divide an event into sub-events and allows a concept of event-headedness which specifies the foregrounded sub-event in the predicate Qualia structure is the mode of explanation composed of formal, constitutive, telic and agentive roles. The lexical-inheritance structure places a lexical word in a broader paradigm and distinguishes it from the others

The advantage of adopting this framework is that it allows one to specify a greater internal structure of a word and provides a richer co-compositional mechanism for creating new words

Sanjukta Ghosh and Anil Thakur are with the Department of Linguistics Banaras Hindu University, Varanasi (e-mail. san subh@yahoo com anilt@bhu ac in)

III. CAUSATIVE COMPOUND VERB CONSTRUCTIONS IN BANGLA AND HINDI

Compound verbs consist of two verbs; the first of them provides the core meaning of the verb and the second modifies its meaning in a significant way. Examples of such verbs from Hindi are cal denaa 'to start', cal paRnaa 'to start suddenly' etc. where the first element is calnaa 'to move' and a broader category or hypernym of the compound verbs constructed. The second verb is responsible for the finer semantic nuisances associated with the compound verb

A causative construction can be defined using the following criteria:

1. It is a single event constituting of two sub-events.
2. These sub-events can be sequential or overlapping
3. One of the sub-events causes the other to happen.
4. There must be some argument sharing between the two sub-events (argument coherence in Pustejovsky's term).
5. Logical entailment relation holds between the two events

A causative construction is generally expressed in Hindi by some morphological markers, viz., adding aa and waa to the verb root. The traditional grammar calls these as first and second causatives respectively Examples of such causatives are from the root /khaanaa 'to eat', /khilaanaa and /khilwaanaa 'to feed and 'to cause to eat (by somebody)'. In Bangla also there is a morphological affix to mark this causativization process, viz., aa , as from /bOl 'to speak' the derived form is /bOlaa 'to make to speak'. However, if somebody makes one to speak with the help of a third person, which is expressed by Hindi waa causative (/bulwaanaa in this case), the choice in Bangla is to use a compound verb like /bolie neoa or /bolie deoa where the first verb is in conjunctive participial form of the verb /bOlaa . The second verb denotes whether the action is for oneself (in case of lenaa/) or for other (in case of denaa/). Bangla exhibits such kind of causative constructions in large numbers. Sometimes they are the only form to express causality. Even in Hindi, such kind of compound verbs are found which express only causality like samjhaa denaa 'to convince'. The following section takes up some of the representative verbs of Bangla and Hindi of this kind and compares them with the corresponding pure morphological causative forms. These compound verb forms have not been mentioned before formally in any work on lexical resources. The present paper suggests a formal model of representation for their meaning differences.

IV. LEXICAL RESOURCE MODEL FOR CAUSATIVE COMPOUND VERBS

This section takes up four major categories of verbal predicates following Vendler (1967) and analyzes one example of the verbs of each category in a lexical resource model adopted in this paper. The four main categories are event, state, accomplishment and achievement predicates. From the first type activity predicate the representative example chosen is Bangla /douRono/ or Hindi /dauRnaa/ 'to run'. Bangla does not have a morphological causative form for this, it uses a complex predicate form of N-V construction dauR korano 'run(n) do-causative-inf' or 'make somebody run' with the causative form of the second verb /kOra For Hindi waa form of this verb, i.e., dauRwaanaa Bangla uses one more verb to make it a more complex predicate. The entry in the lexical model of the verb is illustrated in table I below

Table I

H. *dauRwaanaa*	***H. dauRaanaa***
B. *douR korie neoa douR kOrano*	**B. *dauR kOrano***
EVNTSTR = E1 e1: process	**EVNTSTR = E1 e1: process**
E2 = e2: process	
RESTR = < α	**ARGSTR = Arg1 = animate (x)**
HEAD = e1	**Arg2 = animate (y)**
< = sequential	
ARGSTR = Arg1 of e1 = animate (x)	**Qualia = process-lcp**
Arg2 of e2 = animate (y)	**Agentive = dauRaanaa act (e1,x)**
Arg3 of e2 = animate (y)	
Qualia = cause-lcp	
Formal = dauRnaa (e2,y)	
Agentive = dauRwaanaa act (e1,x)	

The examples below illustrate the difference of the two forms in terms of entailment.

3a. *apni Sobaike douR kOraben manei to ar Sobai doUrobena.*

run do-fut

3b. *aap sabko dauRaeyeMge, iskaa matlab yah nahiiM ki sabhii daureMge hii.*

'If you run everybody, it does not mean that all would run'

4a. **apni amake die procur dour korie niechen kintu ami douRoi ni.*

run do-caus-perf take-pr-prf

4b. **aapne mujhe bahat dauRwaayaa lekin maiN dauRii nahiiM*

'You made me run a lot but I did not run'

The next category is the state class of verbs and the representative example taken here is Hindi samajhnaa 'to understand' and its counterpart in Bangla. For this verb, the second causative form in both the languages is a compound verb. The difference in the semantics between the first and the second causative is also clear from the lexical model given below.

Table II

H. *samajhaa denaa* **B. *bujhie deoa*** **'to cause someone understand'** **EVNTSTR = E1 = e1: process** **E2 = e2: state** **RESTR = < α** **HEAD= e1** **< = sequential** **ARGSTR= Arg1 of e1= animate (x)** **Arg2 of e1= concept (z)** **Arg3 of e2– animate (y)** **Qualia cause-lcp** **Formal samajh (e2,y)** **Agentive=samjhaa-denaa_act (e1,x)**	**H. *samjhaanaa*** **B. *bojhano* 'to make understand'** **EVNTSTR= E1= e1: process** **ARGSTR= Arg1= animate (x)** **Arg2= animate (y)** **Arg3 = concept (z)** **Qualia = process-lcp** **Agentive=samjhaanaa_act (e1,x)** **STATE samajhaa state (e2, y)**

The sentences below capture the difference of the two verbs in both Bangla and Hindi

5a. H. *maiNne usko samjhaayaa lekin wah nahiiN samjhaa*
tried to make understand understood

5b B. *ami oke bojhalam kintu o bujhlo na.*
tried to make understand understood
'I tried to make her understand but she did not understand'

6a. *H. *maiNne usko samjhaa diyaa lekin wah nahiiN samjhii.*
understand gave understood

6b. *B. *ami oke bujhie diechilam kintu o bojhe*
understand-causative-conjunctive participle gave understood
'I made her understand but she didn't understand'

The compound verb of the example 6 expects a state of understanding of the affected argument as a result of a telic event, whereas the verb in 5 does not expect so. It denotes an atelic event.

The next class is the accomplishment verbs where a process event is followed by a state event. The representative example of this class is Hindi /banaanaa 'to make' and its Bangla counterpart gODano . The second causative is again a compound verb in Bangla, Hindi also alternatively uses a compound with the morphological waa causative. The following table illustrates the lexical model for this type of verbs

Table III

H. *banwaanaa/banwaa denaa* **B. *goDie deoa/neoa*** **'to cause some make/create'** **EVNTSTR = E1 = e1: process** **E2 = e2: result** **RESTR = <α** **HEAD= e2** **<α = sequential** **ARGSTR= Arg1 of e1= animate (x)** **Arg2 of e1= material (z** **Arg3 of e2 artistic-creation (y)** **CONST= z** **Formal physical_object** **Qualia cause-lcp** **Formal banaa huaa(e2,y)** **Agentive=banaa-denaa_act (e1,x,z)**	**H. *banaanaa*** **B. *gODAno* 'to make create'** **EVNTSTR= E1= e1: process** **E2= e2: result** **RESTR = <α** **HEAD= e1** **ARGSTR= Arg1= animate (x)** **Arg2= material (y)** **Qualia = process-lcp** **Agentive banaanaa act (e1, x, y)** **FORMAL =creation (e1,y)**

The difference between *banaanaa* and *banwaanaa* and their Bangla counterparts lies in their event-headedness. *Banaanaa* foregrounds its process sub-event whereas *banwaanaa* foregrounds its result sub-event. e.g.

7a. *wah muurti banaataa hai*
He model make-3P-sg
'He makes (clay) models.'

7b. *usne ek din meM yah muurti bannayii.*
He-erg one day-loc this model made
'He made this model in one day.'

The last category of verb in Vendler is the achievement verbs and the representative example here is Hindi pahucanaa 'to reach'. This verb also preferably uses a causative form in both Bangla and Hindi Hindi also has a marginal use of a morphological waa causative. In Bangla, there is no difference between the first and the second causative in formal structure. The lexical model showing the details of the semantic feature of the forms is illustrated in table IV

Table IV

H. ?pahuMcawaanaa/*pahuMcaa denaa* **B. *pouMche dea*** **'to cause something reach'** **EVNTSTR = E1 = e1: process** **E2 = e2: state** **RESTR = <α** **HEAD = e2** **<α sequential** **ARGSTR= Arg1 of e1= animate (x)** **Arg2 of e1= material (y)** **Arg3 of e2= pauhucaa-huaa (y)** **Formal=physical_object/individual** **Qualia cause-lcp** **Formal = pahuMcaa_huaa(e2,y)** **Agentive= banaa-denaa_act (e1,x,z)**	**H. *pahuMcaanaa*** **B. *pouMche dea* 'to make reach'** **EVNTSTR E1 e1: process** **E2 e2: state** **RESTR = <α** **HEAD = e1** **ARGSTR= Arg1= animate (x)** **Arg2 physical object (y)** **Qualia = process-lcp** **Agentive = paucaanaa_act (e1 x, y)** **FORMAL pahucaa huaa (e2 y)**

The basic architecture of this lexical model can be extended to other types of verbs as well including causative and non-causative, simple and complex predicates

V. CONCLUSION

Forming compound verbs to express causativization is a productive strategy in Bangla. We have also shown that though Hindi is believed to have morphological causative only, in many cases they parallel with the compound verbs and in a few cases compounding is the only way of expressing a causative predicate. So far no lexicon has listed the compound verbs with all their semantic nuances and syntactic specifications in the lexicon so as to use this information for further computational operations. In fact, lexicon of this special type can be a rich source for researches in theoretical and computational linguistics. Works in Indian languages within the theoretical framework of Generative Lexicon has begun only in recent years (Raina 2005) and have shown promise for a better analysis of certain linguistic phenomena in Indian languages

REFERENCES

[1] Pustojvsky, J. 1995. *Generative Lexicon*. MIT Press

[2] Achla M. Raina 2005. "Complex Predicates in the Generative Lexicon" Bouillon, Pierrette and Kanzaki, Kyoko (eds.), *Proceedings of GL'2005, Third International Workshop on Generative Approaches to the Lexicon*, School of Translation and Interpretation, University of Geneva, Switzerland, 210-221

[3] Vendler, Zeno 1967. "Verbs and Times". In Z. Vendler, *Linguistics in Philosophy*, pp. 97-121. Ithaca: Cornell University Press.

This paper was presented at LRII-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt of India.

8.7 Handling Polysemous Particles in Multilingual Environment

Anil Thakur and Sanjukta Ghosh

Abstract Particles across languages are tough to process. They exhibit complexity at various levels of processing. They have multiple functional roles based on different syntactic and contextual factors. Traditional grammars have inadequate account of them In this paper, we propose a corpora-based comprehensive analysis of selected polysemous particles, particularly in a multilingual environment where languages are related. The paper also explores the mapping patterns of some of the particles across selected Indian languages and shows the significance of the nature of divergence among them and their NLP implications

Index Terms Particles, Disambiguation, NLP, Multilingual

I. INTRODUCTION

A number of indeclinable words in a natural language are found to be used in multiple grammatical roles (Thakur and Patnaik 2004, Sinha and Thakur 2005). Besides their primary grammatical roles, the other uses of such words are neither easily identifiable nor are they easily categorizable. We treat such "other" functions uses of these words under their particles use For instance, take the Hindi word *Or* 'and'. Its primary grammatical role is to conjoin two words, phrases or clauses sentences, as in (1) below. However, it is also used in other contexts as in (2), where it does not occur in its typical conjunction role

1. *raam **Or** siitaa donoN van gaye.*
   Ram and Sita both forest went
   'Both Ram and Sita went to forest '

2. ***Or** ji, kyaa haal hE?*
   PRT PRT what state is
   'So how are you?'

These uses of *Or* in Hindi, as in (2) where it functions as a sentence starter (SS) particle, are the typical cases of the particle uses of indeclinable words in a natural language. The phenomenon is attested across languages. Among related Indian languages, some examples from Oriya can be illustrative (examples are from Thakur and Patnaik 2004). *je* in Oriya is primarily a complementizer that connects two clauses (the main clause and the subordinate clause), as in (3a). However, it can also occur in other atypical uses as in other examples in (3) below

3.a. *mu jaane **je** se kaalii aasibe.*
   I know COMP he tomorrow come-FU
   'I know that he will come tomorrow '

b. ***je** tume mo kathaare ete raagi jaucha!*
   PRT you my words so much angry getting
   'Oh you are getting so angry with what I'm saying!'

c. *mu tumaa kathaa kaahinki sunibi **je**!*
   I your words why listen-FU PRT
   'Why should I listen to you?'

In (3b), *je* occurs as a sentence starter particle whereas in (3c), it occurs as a pragmatic particle. In traditional dictionary, however, *Or* or *je* would be entered only in their primary role grammatical category, that is, that of conjunction and complementizer respectively. As the discussion above clearly shows, the uses of such words in multiple senses need to be identified and separately categorized according to the different roles they occur with. Particularly, in the context of applied natural language processing, the task of identification and classification can play an important role. Since there are lots of commonalities across related Indian languages, it would be appropriate to examine the issue in a multilingual perspective. Our assumption is that the analysis of the nature of a particular particle in an Indian language can throw much light on the counterparts of those particles in other related Indian languages. With this assumption in mind, we examine the polysemous nature of a couple of particles and their distribution across Hindi, Bangla and Maithili. On the basis of our analysis, we propose that particles in a multilingual environment need to be examined with the help of large corpora and they need to be identified and categorized for their use in multiple roles. Some of the immediate advantages of such study include capturing the divergence patterns across related Indian languages (particularly for machine translation among them) and for tagging purpose for both monolingual and parallel corpora. The relevance of the study in the context of scientific and technical terminology can be seen in the light of the fact that there is much gap between availability and requirement of technical terms (particularly in Indian languages) to define and describe various concepts in many of the newly emerging branches of linguistics such as computational linguistics, neurolinguistics, cognitive linguistics, etc. The paper proposes to provide a ground for augmentation and standardization of scientific and technical terms, particularly in the area of linguistics.

Anil Thakur and Sanjukta Ghosh are with the Department of Linguistics Banaras Hindu University Varanasi (e-mail: anilt@bhu.ac.in, san subh@yahoo.com)

II. SOME PARTICLES AND THEIR DISTRIBUTION

A. *Some particle-uses of 'ki' in Hindi, Bangla and Maithili*

ki Alternative-Question Particle

In many Indian languages such as Bangla, Maithili, Bhojpuri, etc. *ki* is a question particle (its Hindi counterpart is kyaa). However, in Hindi too, besides in Bangla and Maithili, *ki* occurs in certain cases in the sense of an alternative-question particle. Examples in (4) show distribution of *ki* across Hindi, Bangla and Maithili[1]

4. a. mujhe nahiiN maalum ki raam aa-yaa (yaa) ki nahiiN?
   to-me not known COMP Ram came or PRT not
   b. aamii jaani na je raam aaSbe ki (aaSbe) naa?
   I know not COMP Ram come-FU PRT come not
   c. hamaraa nai bujhal achi ki raam Otaa ki nai (Otaa)?
   to-me not known COMP Ram come-FU PRT not come-FU
   'I do not know whether Ram comes or not?'

As we notice, the first use of *ki* in Hindi and Maithili is that of complementizer (COMP) (the Bangla counterpart is *je*). The second use of *ki* in Hindi and Maithili and the only *ki* in the Bangla sentence is used to mark alternative-question. The sentences of this type can be used either to ask question or to reply in a normal discourse. Since this *ki* is used to indicate question in the alternative possibility of answer in positive or negative, we can call this *ki* alternative-question particle.

ki Purpose-Clause Particle

ki in Hindi and Maithili can also be used to indicate a purpose clause function. Normally purpose clause in Hindi and Maithili is marked by *taaki* 'so that' and *jeki* respectively. In Bangla, it is marked by *jate* 'so that'.

5 a. jaraa jor boliye ki ham bhii sune.
please loud speak PRT we EMP listen
b. kTu jore bolun jate aamraao suni.
please loud speak PRT we-EMP listen
c kani jorsa baaju je/ki hamhusab suni.
please loud speak PRT we-EMP listen
'Please speak (a bit) louder so that we may also listen'

[1] The (a), (b) and (c) examples onwards represent Hindi, Bangla and Maithili respectively.

We notice that both Hindi and Maithili use *ki* for this function, Bangla does not show this option, in which only *jaate* can be used.

ki. Correlative-Clause Particle

ki is also optionally used to mark correlative clause along with the correlative words (such as *jitanaa*, *jahaaN*, *jo*, *jab*, etc.) in both Hindi and Maithili However, in Bangla this use of *ki* is not attested, as the examples in (6) illustrate. The optional use of *naaki* in Bangla is highly marked

6. a. mEN utanaa nahiiN jaanataa jitanaa ki aap jaanate hEN
   I that much not know as much PRT you know AUX
   b. aami OtoTaa jaanii naa jOtaTaa (?naaki) aapanii jaanen
   I that much know not as much PRT you know AUX
   c. ham otek nai jaanai chii jatek ki aahaaN jaanai chii
   I that much not know aux as much PRT you know AUX
   'I do not know as much as you know.'

ki: Wish/Possibility-Clause Particle

ki is also used to introduce a clause to the main clause of wish-possibility type, as in (7-8). The Bangla counterpart of this *ki* is *je* and *jodi* in (7b) and (8b) respectively. Maithili[2] uses either *ki* or *je* or can also use both together in some cases.

7. a. kahiiN yEsaa na ho ki aap gir paRe
   may be PRT you fall down
   b. jEno Emon na hoy je apnii poRe gelen
   may be PRT you fall down
   c. kahiiN i na hoye je/ki ahaaN gir paRii
   may be PRT you fall down
   '(Take care) may be, you fall down.'
8 a acchaa hotaa ki ham pahale aaye hote
good happened PRT we earlier come aux
b. bhaalo hoto jodi aamra aage aastam.
good happened if we earlier come
c. niik hoit je/ki ham pahine aayl rahitauN.
good happened PRT we earlier come aux
'It could have been nice if we had come earlier.'

In (8), Bangla uses only *jadi* 'if' whereas in Hindi and Maithili *ki* and *je/ki* respectively can be used to indicate this particular clause construction

[2] The use of ki in Maithili may be because of the influence of Hindi.

ki Temporal-Clause Particle

Both Hindi and Maithili have the temporal adverbial use of *ki*. Bangla, however, does not have such use. In Bangla, the identical construction is expressed by the exclusive focus suffix, the counterpart of hi in Hindi. The similar use of hi is also attested in Hindi

9. a. raam jEse hii ghar pahuNcaa ki baarish band ho gayii.
   Ram as soon as home reached PRT rain stopped be went
   b. raam baaRii pouchono matroi pouchotei brisTi bOndho hoye gElo
   Ram home reaching as soon as rain stop be went
   c. raam jahinaa ghOr pahuNcalaa ki brisTi band bhO gelai
   Ram as soon as home reached PRT rain stop be went
   'As soon as Ram reached home, it stopped raining.'

The particle-uses of *ki* shown in these examples are only illustrative of the point discussed in this paper. A detailed study of the multiple functions of *ki* along with their mapping patterns (from the point of view of Hindi-English machine translation) is given in Sinha and Thakur (2005). Besides *ki*, many other words can be shown to have this property of occurring in multiple functional roles. We can take *kyaa* in Hindi and its Bangla and Maithili counterpart *ki* to illustrate the point.

B. *Some particle-uses of 'kyaa ki' in Hindi, Bangla and Maithili*

kyaa/ki is primarily an interrogative pronoun as well as a question particle (for yes-no question) (a detailed study on the disambiguation of *kyaa* in Hindi is given in Sinha and Thakur forthcoming) in these languages. However, *kyaa/ki* also occurs in other functional roles in these languages, a few of which can be shown in this section

kyaa/ki Emphatic Assertion Particle

kyaa and *ki* are used to indicate emphatic assertion in Hindi and Maithili respectively. However, in Bangla, this function is marked by a topic particle to as shown in (10) Hindi and Maithili also use topic particle *to* and *ta* respectively for topic marking

10. a. yah kyaa paRii hE tumahaarii kitaab !
    this PRT lie.PRF be.PR your book
    b. ei to poRe roeche tomar boi !
    this PRT lie PR PRF your book
    c. i ki (raakhal) chau tohar bOhi
    this PRT kept is your book
    'Your book is lying very much here.'

kyaa ki Emphatic Negation Particle

kyaa is also used to indicate emphatic negation and its Bangla and Maithili counterparts *ki* is used for the similar function in these languages, as in (11) below.

11. a. vah tumhaaraa kyaa bigaar legaa
    he your PRT harm do FU
    b. o tomar ki khoti korbe. (B)
    s/he your PRT harm do FU
    c. se tohar ki bigaarletau. (M)
    s/he your PRT harm.FU
    'He cannot do any harm to you.'

kyaa ki. Temporal Adverbial Particle

Another use of *kyaa* in Hindi and its Bangla and Maithili counterpart ki is to mark temporal adverbial, as in (12).

12. a. vah dikhaa kyaa ki sab uskii or lapak paRe.
    he seen PRT that all his towards jump fell
    b. oke dekhlo ki SObai or opor jhaNpiye poRlo.
    her/him saw PRT all her/his upon jump.CP fell
    c. o dekhi ki paRal ki sOb okare dish lapaki paRal
    he seen PRT happened that all his direction jump fell
    'The moment he was seen, all jumped towards him '

However, we may also notice that in Bangla, the syntactic environment of this use of *ki* is different from that of *kyaa* in Hindi and *ki* in Maithili. In the discussion above, we have listed a couple of particles and some of their multiple uses in Hindi, Bangla and Maithili. This has been done with a view to showing the state of indeterminacy about the different functions of such words and their appropriate categorization.

III. IDENTIFICATION AND CATEGORIZATION

As we have seen from the discussion of only two particles, Hindi and its Bangla and Maithili counterparts have much in common with respect to their

distributional behavior. However, we have also noticed some of the differences that they have. Both their common features and those on which they differ can be used to identify and categorize the behavior of the particles in general. Secondly, the identification of their commonalities and differences is useful for building comparative lexical resources on particles across these languages. As we can notice, the identification of the different roles of such particles is dependent on both their syntactic and semantic analysis such as the position of their occurrence in the sentence, the nature/type of the sentence, the occurrence of some other element in the sentence, etc (Sinha and Thakur 2005). Thus, we notice significant commonalities across languages with respect to the identification strategies, too. On the basis of these observations, we can propose that a study for identification and categorization of the polysemous words/particles can be more fruitful in multilingual environment.

IV. DIVERGENCE AND NLP APPLICATIONS

It is known that a study of structural similarities as well as divergence patterns among the related Indian languages will go a long way in obtaining mapping patterns across these languages. For instance, for the purpose of machine translation across Indian languages a study of this kind can be quite indispensable. Another advantage of such study can be seen with respect to the enrichment of parallel lexical resources that can be useful for various natural language applications. Exact identification and categorization of particles, for example, will lead to appropriate tagging. For various NLP applications in Indian languages (e.g., Hindi, Bangla, Maithili), a contrastive analysis of the languages is important.

REFERENCES

[1] Patnaik, B. N. (1979) "Syntax and Semantics of the Indeclinable *bohi* in Oriya." *Language Forum 5.1.*

[2] Sinha, R M K. and Thakur, Anil (2005). "Handling *ki* in Hindi for Hindi-English MT". *Proceedings of MT SummitX*, Thailand.

[3] Sinha, R M K. and Thakur, Anil (forthcoming) "Disambiguation of *kyaa* in Hindi for Hindi to English Machine Translation". *Indian Linguistics.*

[4] Thakur, Anil and B N. Patnaik (2004). "Indeclinables and Natural Language Processing". *Proceeding of the International Symposium on MT, NLP and Translation Support Systems (iSTRANS)*, New Delhi: Tata McGraw-Hill Publishing Co.

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.8 Morphological Analyzer for Great Andamanese Verbs: Implementing a Concatenative Template

Narayan K. Choudhary, Anvita Abbi and Girish Nath Jha, Jawaharlal Nehru University, New Delhi

Abstract This paper presents an account of the verb phrase morphology of Great Andamanese, an endangered language of the Andaman Islands. The paper is based on the research work done in the Andaman Islands among the people of the endangered tribe The verb phrase in Great Andamanese takes as constituents the morphemes carrying the content of causatives, subject and object clitic, negative, prohibitive negative, class marking consonant or thematic consonant (Abbi 2003, 2006) and TAM markings. All of these features are affixed to the verb root or lexeme -the only obligatory element in the verb phrase. An illustrative schema to the Great Andamanese verb phrase can be given like the following

C A U S - S B J . C L - O B J . C L - R E F L - CAUS/NEG→VR VL←NEG-CLSM-TAM

The schema works on constraints at the affixal level which include order, optionality and obligatoriness. The morphophonemic rules such as epenthesis, vowel deletion, assimilation that operate in the varying forms of verb phrase are not discussed here.Using a lexicon based approach to develop a morphological analyzer for the verb phrase in Great Andamanese, the paper presents the mechanisms used in developing a program that analyzes the verb phrase given the Great Andamanese text as input.

Index Terms Concatenative Morphological Template, Great Andamanese, Natural Language Processing, Verb Morphology

I. INTRODUCTION

THE Andaman Islands are a group of more than 500 islands situated in the Bay of Bengal. It is inhabited by a community that has been living there for long, in complete isolation. The earliest record of these people belonging to the Negrito stock (Hagelberg et. al., 2003, etc.) is found in, among others, Ptolemy (2nd C. AD), I-Tsing (672 AD) and Marco Polo (14th C. AD).

Among the four primitive *tribes* the Great Andamanese, the Jarwas, the Onges and the Sentenelese - of the Andaman Islands, the Great Andamanese, till a hundred years back, were the most populated and influential people

The linguistic study of the rapidly vanishing voices of the Great Andamanese can be said to start with M.V. Portman's *Manual* in 1887 followed by other major works like that of E.H. Man's *Dictionary* (1919), Manoharan (1989) and Abbi (2001, 2006)

The Great Andamanese is a cover term assigned to a conglomerate of the ten tribes most of whom succumbed to the colonial pressure that started with the British and is still continuing in its new avatar. The present population (around 50) is dominated by the Jeru tribe with a few speakers (around 7) of the language. As the new generation is reluctant to learn the language of their forefathers, the language is under an imminent threat of extinction. Great Andamanese is an unwritten tribal language. The data presented in this paper is drawn from first-hand data elicitation in the field.

II. UNRAVELING THE VERB PARADIGM SCHEMA OF GREAT ANDAMANESE

Great Andamanese is an agglutinating language and is of the SOV type, meaning thereby it is a verb final language. The verb phrase of the language is a complex entity constituted of several grammatical morphemes. A verb root in a verb phrase is preceded by several prefixes as well as followed by two or more suffixes. These prefixes and suffixes encode several grammatical functions such as subject and object information, various modalities such as negation and mood. In addition, tense marking is suffixed to the verb stem. In all, the possibility of various types of affixation to the verb root or lexeme can be illustrated using the following schema (see Fig. 1).

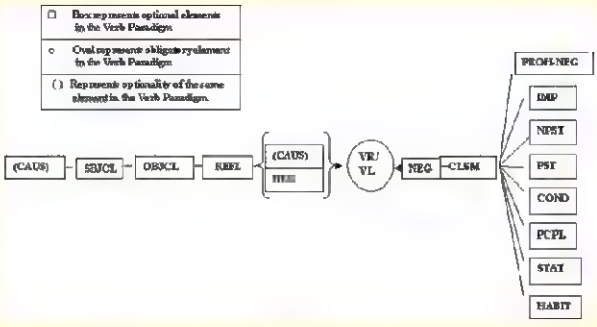


Figure 1. Verb Schema of Great Andamanese

Narayan Kumar Choudhary and Anvita Abbi are with the Centre for Linguistics, Jawaharlal Nehru University, New Delhi (e-mail. choudhary narayan@rediffmail.com, anvitaabbi@gmail.com)

Girish Nath Jha is with the Special Centre for Sanskrit Studies Jawaharlal Nehru University New Delhi (e-mail. girish@gmail.com)

For example :-

i	t^{h}utconnep^{h}obe
	t^{h}ut-conne-p^{h}o-b-e
	1SG.CL go NEG CLS IND
	I do not go.

ii	ut^{h}uncikamo
	u-t^{h}u-n-ci-k-amo
	3SG.CL-1SG.CL-REFL-comes-CLS-COND
	If he comes to me

There are at most five morphemes that can possibly be prefixed to the Verb Root (VR) or verb lexeme (VR) while at most three morphemes that can be suffixed to it. The only obligatory element in the verb phrase (VP) is the VR or VL. Thus a verb phrase with maximum number of affixes will have the structure as the following-

CAUS-SBJ CL-OBJ CL-REFL-NEG→VR←CLSM-TAM
Or,
CAUS-SBJ CL-OBJ CL-REFL→VR←NEG-CLSM-TAM
Or,
SBJ CL-OBJ CL-REFL-CAU→SVR←NEG- CLSM-TAM

For example we have verb phrases like 'p^{h}ute∫amo' and 't^{h}uŋolobom' as in example sentence number iii below, 'ut^{h}uncikom' as in iv and 'ŋutuncek^{h}o' as in v.

iii ŋut^{n}i mit^{h}aibi te∫e p^{h}ute∫amo t^{h}oŋolobom
ŋu-t^{n}i mit^{h}ai-bi te∫-e p^{h}u-te∫-amo t^{h}o-ŋol-o-b-om
2SG-1SG OBJ sweet-ACC give-IMP NEG-give-COND
1SG CL-cry-UPV-CLS-NPST
If you do not give me the sweets I will cry.

iv cya:k ocikom kɔil to ut^{h}uncikom
cya-k o-ci-k-om kɔil to u-t^{h}u-inci-k-om
what-DIRTC 3SG.SBJ.CL -come-CLS-NPST later
EMPH 3SG.SBJ.CL-1SG.OBJ.CL-come-CLS-NPST
Where will he go, later he will come only to me.

v ca:y k^{h}udi ŋutuncek^{h}o
ca:y k^{h}udi ŋu-tun-cek^{h}-o
what for 2SG.SBJ.CL-REFL -angry-PST
Why did you get angry?

III. A FRAMEWORK FOR THE ANALYZER

The Great Andamanese Verb Analyzer (GAVA) is a five module program that takes Great Andamanese text as input, in IPA (*using Lucida Sans Unicode or Arial Unicode MS fonts* and analyzes the verb phrases in it. The five modules are in fact the five processes that the input text undergoes. This has been illustrated in the following diagram (see Fig. 2).

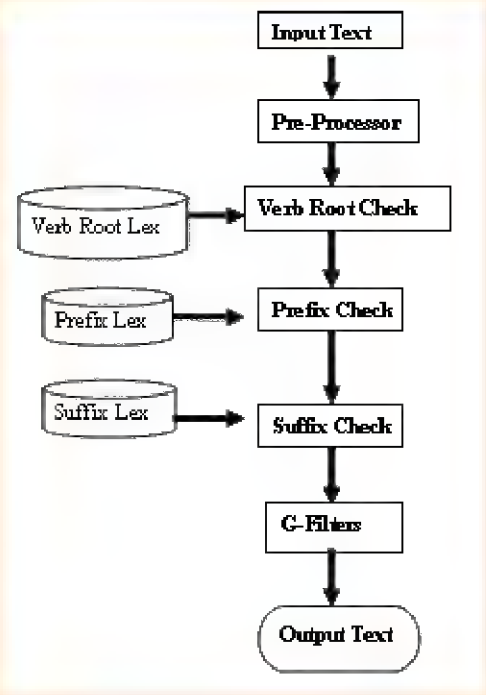


Figure2. A Model Diagram of GAVA

The pre-processor module first filters the input and checks whether any unwanted elements are there in the input text or not. If this is the case, it either corrects the input or leaves it as it is for the consideration of user.

The verb root module searches for the verb roots or lexeme in the input text and segments them from the string. The remaining part of the input string is sent for further analysis in the next modules.

The prefix module takes the elements that are to the left of the verb lexeme and analyzes them by matching each of the possible strings with the prefixes in the prefix lexicon and stores the results for display

The suffix module takes the elements that are to the right of the verb root or lexeme and analyzes them matching each of the possible strings with the suffixes in the suffix lexicon and stores the results for display.

The G-Filters module is the last module that implements the grammatical rules. If the system has not found the right analysis of the input text or if there is some ambiguity or violation of some rules, these are checked through rules here.

The final result is displayed as Unicode HTML on a JSP web front.

A. *POS tags for Great Andamanese verbs*

Each of the verbs have been tagged with their meaning and an additional identifier of VR in the lexicon. No classification of the verbs as per transitive intransitive or on any other criteria has been made. The linguistic resources used have all been prepared on the basis of first- hand data collected by Abbi (2001, 2005) and Choudhary (2005-06) and compared with available other printed forms. The program uses lexicon that is basically text files of small sizes

B. *Tagged Lexicon*

There are three types of grammatical categories that are used for the Great Andamanese Verb Analyzer.

Verb Roots

Prefixes

Suffixes

All these categories are tagged properly. The prefixes and suffixes have an additional tag of PREF and SUFF respectively. This is for specific use of the program and is also displayed in the output text.

The lexicon of verb lexemes[1] contains about 120 verb roots and non-verb roots. All are verbal lexemes. The longest verb lexeme in Great Andamanese is of three syllables containing eight characters. The frequency of monosyllabic roots is higher than disyllabic roots and that of latter is much higher than tri-syllabic roots. All these roots have been arranged in the lexicon in an ascending order of the number of characters present in the lexeme to facilitate better search by the program.

There are a total of 52 prefixes and 20 suffixes at present. The number of affixes has grown up because there are allomorphic variations. Thus a morpheme with a gloss of 1SG.SBJ.CL has 6 variations, 1SG OBJ CL has 4 variations etc. The following table gives a list of variations in the clitics attached to verbs. A clitic is a morpheme that has the syntactic characteristics of a word but shows evidence of being phonologically bound to another word. The clitics given in table-1[2]

TABLE I
A list of pronominal verbal clitics in Great Andamanese

Name of the Prefix	No of Variants	Variant Forms
1SG.SBJ.CL	6	t̪ʰu, t̪ʰa, t̪ʰo, t̪ʰe, t̪ʰ, t̪ʰut
1PL.SBJ.CL	2	me, mut
1SG OBJ CL	4	t̪ʰu, t̪ʰa, t̪ʰi, t̪ʰe,
2SG.SBJ.CL	7	ŋu, ŋa, ŋe, ŋɛ, ŋi, ŋ, ŋut
2SG OBJ CL	4	ŋu, ŋa, ŋe, ŋɛ, ŋí, ŋut
3SG.SBJ.CL	7	a, o, a, e, aka, uku, du.
3SG.OBJ CL	8	aka, ek, ɛk, ek, ik, it, ut, et, i
3PL.SBJ.CL	2	tu, n
3PL.OBJ.CL	NA	Not Available

(The first person and second person clitics are homophonous in subject and object position while third person clitics are not)

C. *Rules*

There is only one rule implemented. This rule takes care of the ordering problem of the prefixes that emerges due to identical forms of the clitics. For example, t̪ʰu can be used as both subject clitic and as object clitic. Similarly there are other clitics that have homophonous entries. This problem can be solved by the constraint of ordering. In a verb phrase there can be no more than two clitics. As the language is of SOV nature, the subject clitic precedes the object clitic no matter what the phonetic shape is.

If a single clitic in a verb phrase is found, it is assigned the tag of subject clitic. However, it is not always necessary that the single clitic in the VP is subject clitic. If the subject is omitted or is not a pronominal category in the verb phrase, it can be an object clitic in case of transitive verb root. In this case, the solution lies only in the context of the whole clause.

There are also morphophonemic changes involved in the verb morphology of Great Andamanese, which will constitute the subject matter of the next paper.

D. *Implementation Strategies*

The program has been prepared on a Windows platform with tools and techniques as described below. This program however is platform independent and can run on any platform.

1 Verbal lexeme in Great Andamanese consists of minimum of a verb root in case of verb intransitive and maximum of two argument markers prefixed to the verb root in case of Verb transitive. Thus implies that transitive verb root is always prefixed by an optional subject and obligatory object clitic to gain a lexemic status to take part in the verb analyzer.

2 The list is not final as it is based on a limited source of data. There may be more or less variants, their names and forms. More specific study on this topic is warranted.

IV. AN OVERVIEW OF THE TOOLS AND TECHNIQUES USED

Table II gives an overview of the tools and techniques used in developing the program.

TABLE II
An Overview of the Tools and Techniques Used

I. Front end
   A. JSP, HTML, CSS, Java Script
II. Java Objects
   A. Pre-processor
   B. Analyzer
      1) Search Parts()
         a) Verb root
         b) Prefixes
         c) Suffixes
      2) gFilter()
      3) reorder()
III. Back-end
   A. Data files stored in UTF-8
IV. Webserver
   A. Apache-Tomcat

A. *Front End*

At the front end of the program the technologies like the JSP, HTML, CSS, Java Script have been used. The following is a brief introduction to these technologies and how they have been implemented in developing the program

The front end opens in a web browser that is based locally on the user's computer.

1) Java Server Pages

The java server page used here utilizes all of the four items discussed above. It uses first, the html coding convention and initializes the style sheet, the java objects from AVT*agger.class* as servlets.

Using small Java programs (called "Applets"), web pages can include functions such as animation's, calculators, and other fancy tricks. Java programs are of three kinds

Stand-alone executable programs
Applets
Servlets

a) Cascading Style Sheets

Here the CSS has been used to bring text in a particular font namely the Lucida Sans Unicode Another font named Arial Unicode MS can also be used for the purpose of entering the input text in Great Andamanese

b) HTML

HTML or Hyper Text Mark-up Language is the base of the front end of the interface on which other objects namely that of Java Objects and CSS has been embedded

B. *Java Objects*

The JSP file called the *andverbs.jsp* uses a java object called AVTagger which uses the services of Pre-processor The *Pre-processor* object filters the input text and checks whether the input text is a potential Great Andamanese text or not. The *AVTagger* object is in fact the analyzer program that processes the input text as rendered by the pre-processor.

As described briefly above, there are five modules of the GAVA program. Below given is the description of each of the modules

1) Pre-processor

The input first goes to the pre-processor module and checks whether any undesired elements such as punctuation marks or other control characters, numbers etc. are not given in input. If this is the case, either it corrects the input text itself or removes them from going into further analysis.

2) Analyzer

AVTagger is the file that is the most important to the program. Two Java APIs from the Java library have been imported to be used in this object

The analyzer uses several functions and methods to analyze the GA verb.

a) parseVerbs

This is the main calling function which gets all the work done by using services of other functions methods. This function first gets the pre-processing done on the entire text. Then it tokenizes the output of the pre-processor based on space character. Then by calling the search Parts() function, it processes each word for verb, and affixes (to a maximum of 5 prefixes and 3 suffixes)

b) Search Parts

The search is then for the parts starting from the whole of the input to the last available string in the input text until the search is complete or there are no characters left to be searched and matched with the lexicon (or first five prefixes and first three suffixes have been searched). The search is processed in three modules. The search Parts module assumes the role of searching verbs, prefixes or suffixes when an appropriate call is made for each kind of search

c) G-Filter

It is here that the grammatical rules not covered in the previous modules are taken care of. The rules that are applied can be classified broadly into three categories, namely, reordering, constraints and recursivities

(1) Reordering

The same key may have more than one value. There are pronominal clitics that have identical shapes as subject and object clitics. In this case, simple search results in a random choice may be wrong. To bring surety of the results, some rules have been drawn.

(a) Ordering of the Segmented Items

Meta Rule: Follow the ordering rule as prescribed in the verb paradigm. Take the order as given in the input string

(b) Clitics Reordering

For the clitics having homophonous forms (e.g. 1st and 2nd person clitics), the following rules apply:

Rule A. If there is only one clitic preceding the verb root, take it to be Sbj cl by default

Rule B. If there are two clitics preceding the verb root, take the first one as Sbj cl and the second one as Obj cl

(2) Constraints

The input verb phrase in Great Andamanese has a limited number of prefixes and suffixes. These numbers work as constraints and the system would not recognize the input if it has more than the required number of affixes.

(3) Recursivity

As there may be systemic ambiguities regarding the verb roots or the prefixes after the first round of processing of the input text, to handle this, the options multiple values are again sent back for better results.

As there may be more than one affix in the input word, the system must analyze all of this, one by one For this, the system must be recursive to search for different affixes in the same lexicon

C. Back-End: Data files stored in UTF-8

The GAVA uses data files of three types of lexicon as described above. These are annotated lexicon of verb roots, prefixes and suffixes.

D. Webserver: Apache Tomcat 4.0

We have used Apache Tomcat technology for the web server.

V. EVALUATING THE PROGRAM

After successful testing of the verb phrases of Great Andamanese, more than 90% results were found correct. The verb types may be divided on the following basis: 1. Number of prefixes and suffixes 2. Types of Verb Roots based on the number of characters or syllables

So far, I have tested a list of verb phrases extracted from a set of model sentences containing a total of 129 verb phrases (Choudhary, 2006), with a satisfying correct result of 94%

VI. CONCLUSION

As the ambition of this project is to develop a computational framework for the verb morphology of the language, the GAVA program does not aspire to account for an exhaustive list of the verb roots and lexemes in the language under discussion. It uses a list of about 130 verb lexemes. It is basically a morphological analyzer. It is highly scalable and portable system. As an NLP program, it can be used in several ways. It can serve as a template for further work on computing of this language or other languages having morphological systems. As the system developed is highly scalable, it can be easily adapted and extended to suit the needs of other languages as well

GAVA can also serve as a subsystem for major NLP systems on this language or other languages with like structures. The major programs may be a general purpose parser, machine translation systems, speech recognition systems, corpus anralyzers etc.

ABBREVIATIONS USED

1 First Person 2 Second Person 3 Third Person Arg Argument Marker AUX Auxiliary CAUS Causative CL Clitic CLS Class Marker Consonant or Thematic Consonant COND Conditional EPV Epenthetic Vowel EXCL Exclusive EXIST Existential Gen genitive HABIT Habitual IMP Imperative INCL Inclusive IND Indicative NEG Negative NPST Non-Past OBJ Object PCPL Participle PL Plural PREF Prefix PST Past PROH.NEG Prohibitive Negative REFL Reflexive SG Singular STAT Stative SBJ Subject SUFF Suffix VL Verb Lexeme VR Verb Root

APPENDIX

Lexicon A: The Verb Roots and lexemes <verbroots txt>

emp^{h}orol = turn_VR

kaŋyɔrɔ = come_frequently_VR

kaŋɔrɔ = come_frequently_VR

ereŋk^{h}ol = play_VR

ravu fro = winnow_VR

ekterʈɔ = throw_VR

untɛle = call_with_happiness_VR

emp^{h}il = die_VR

bɔk^{h}um = know_(neg)_VR

tabiŋo = think_VR

aratta = convince_VR

ekak^{h}u = open_VR

embele = overflow_VR

akaile = return_VR

tɛrtʰu = take_out_VR
raliʃo = finish_VR
bɔrɔtʰ = fall_VR
ɛrence = fight_VR
conne = go_VR
cɔnne = go_VR
rɛpʰo = climb_tree_VR
erŋol = write_VR
itpʰu = cut_VR
tɛrta = tell_VR
utlub = open_VR
mekʰu = bloom_VR
birəŋ = redden_VR
tɛbol = run_away_VR
ertedʟ = see_VR
raʃui = cook_VR
beliŋ = cut_VR
elukʰ = pick_(caus)_VR
ʈʰibi = live_VR
bereŋ = pour_VR
ʃerep = cut_VR
rapʰo = cut_VR
ʈʰulu = kick_VR
kʰole = laugh_VR
ekter = push_VR
ipʰil = throw_VR
ɛʃilo = shake_VR
ka:ra = rise_VR
tɛrto = shoot_arrow_VR
batʰe = slap_VR
rokʰo = ready,_to_get_VR
bilup = remember_VR
boʃuʈ = hit_VR
olam = tire_VR
tʰud = pierce_VR
belɔ = aux-clsm-pst_VR
boʃo = beat_VR
eban = make_VR
biŋo = hear_VR
dʟuoc = hear_VR
eule = see_VR
meli = return_VR

bitʰ = sink_VR
jiyo = stay ebb AUX_EXIST VR
koin = wake_up_VR
cɛkʰ = to_be_angry_VR
ʈɔpʰ = bathe_VR
ʃuɲe = blow,_of_nose_VR
ʈɔl = break(intr.)_VR
undʟu = break_VR
buli = defecate_VR
juvu = fly_VR
emfe = jump_VR
inci = go_VR
tɔle = mix_VR
ralɛ = moonset_VR
bele = overflow_VR
ʈɛnɔ = pull_VR
cokʰ = row_VR
koʈɛ = serve_food_VR
ʃimu = soak_VR
buli = take_away_VR
cɔpʰ = to_be_enough_VR
beno = sleep_VR
jira = speak_VR
ʈoya = stand_up_VR
kɛle = stay_VR
lele = swing_VR
emaʈ = run_VR
coŋ = get find_VR
cɔŋ = get find_VR
ʃɔr = sing_VR
noe = knit_VR
boi = ask_VR
bɔi = ask_VR
eɲo = come_VR
tʰu = come_out_VR
ŋol = cry_VR
ŋɔl = cry_V
caʈ = do_VR
bɔl = peel_VR
tɔl = roam_around_VR
eul = see_VR
iye = catch_VR

tokʰ = close_VR
ʃui = cook/burn_VR
kaɲ = touch_VR
buʈʰ = fall_VR
iji = eat_VR
tɛʃ = give_VR
ʃol = walk/hang_VR
mok = leave_VR
muk = leave_VR
ɲyo = live_(home)_VR
rɔʃ = love_VR
oɖu = paste_VR
kʰi = pour_VR
kʰu = drink_VR
cɛr = rain_VR
bor = scratch_VR
lɛb = sweep_VR
cɔk = do_well_VR
ʃiʈ = hunt_VR
lub = pluck_VR
uno = sit_down_VR
ʈob = steal_VR
ɲɔʈ = swim_VR
ʃir = wash_VR
ɲa = bark_VR
ku = burn_VR
ɲa = eat_VR
cu = have_VR
ɖe = shut_up_VR
eb = take_VR
co = tie_VR
ie = give_VR

REFERENCES

[1] Abbi, Anvita 2003. *Vanishing Voices of the Languages of the Andaman Islands* Paper presented at the Max Planck Institute, Leipzig

[2] Abbi, Anvita. 2005. *Is Andamanese Typologically Divergent from Standard Average Andamanese*. In the 6th Biennial Meeting of Association for Linguistic Typology Padang, West Sumatra, Indonesia 21-25 July

[3] Abbi, A. 2006 *Endangered Languages of the Andaman Islands* Lincom-Europa Munich.

[4] Choudhary, Narayan K. 2006. *Developing a Computational Framework for the Verb Morphology of Great Andamanese*. Unpublished Dissertation, Jawaharlal Nehru University, New Delhi

[5] Endicott, Phillip, M. Thomas, P. Gilbert, Ch. Stringer, C. Lalueza-Fox, E. Willerslev, A.J. Hansen, A. Cooper. 'The Genetic Origins of the Andaman Islanders' The American Journal of Human Genetics. No. 72 (1), January 2003 Report no 178

[6] Hagelberg, Erika, Lalji Singh, K. Thangaraj, A.G. Reddy, V.R. Rao, S.C. Sehgal, P.A Underhill, M Pierson, I.G. Frame. 'Genetic Affinities of the Andaman Islanders. A Vanishing Human Population. *Current Biology*, January 21, 2003 13, pp 86-93

[7] Man, E.H. 1919. *A Dictionary of the South Andaman Language*, Indian Antiquary

[8] Manoharan, S. 1989. *A Descriptive and Comparative Study of the Andamanese language*. Anthropological Survey of India Calcutta.

[9] Portman, M V [1898] 1992 (reprint). *Manual of the Andamanese Languages* Manas Publications Delhi

[10] Radcliffe-Brown, A R. 1948 *The Andaman Islanders*. Free Press Illinois

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt of India.

8.9 UNL Punjabi Deconverter

Sandeep Singh Spall and Parteek Bhatia, Thapar Institute of Engineering and Technology, Patiala, Punjab

Abstract— This paper discusses the Interlingua approach to machine translation. Here Universal Networking Language (UNL) has been used as the intermediate Language. In this paper the Deconverter from UNL to Punjabi language has been described. The information needed to generate the Punjabi sentence is available at different linguistic levels. The process of deconversion involves case marker generation, morphology phase and syntax planning phase.

I. INTRODUCTION

THE deconverter is a language independent generator that provides a framework for syntactic and morphological generation as well as co-occurrence-based word selection for natural collocation It can deconvert UNL expressions into a variety of native languages, using a number of linguistic data such as Word Dictionary, Grammatical Rules and Co-occurrence Dictionary of each language

It should work for any language by simply adapting a different set of the grammatical rules and Word Dictionary of a language. For this purpose, the function of DeConverter should be powerful enough to deal with a variety of natural languages but never depend on any specific languages. As a result, the Deconversion capability of Deconverter covers context-free languages, as well as context-sensitive languages.

First of all, Deconverter transforms the sentence represented by an UNL expression - that is, a set of binary relations - into the directed hyper graph structure called **Node-net** The root node of a **Node-net** is called **Entry Node** and represents the main predicate of the sentence. It then applies generation rules to every node in the Node-net respectively, and generates the word list in the target language. In this process, the syntactic structure is determined by applying Syntactic Rules, while morphemes are generated by applying Morphological Rules.

The Deconverter can be logically portioned into three phases as

1) Case marking phase
2) Morphology phase
3) Syntax planning phase

II. THE UNL PUNJABI DECONVERTER STRUCTURE

Deconverter has been designed by UNU/IAS as a language independent generator that provides synchronously a framework for morphological and syntactic generation and word selection for natural collocation. The structure of Deconverter is given below

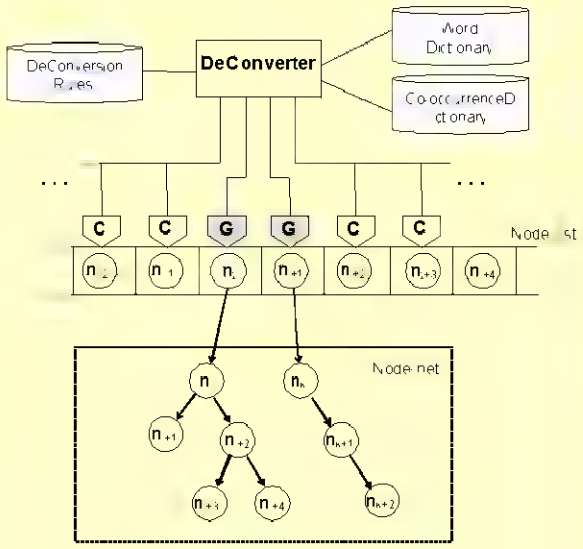


Figure 1.1: Structure of DeConverter

Here, "G' indicates a Generation Window; "C" indicates a Condition Window

Deconverter operates on the nodes of the **Node-list,** and inserts nodes from the **Node-net** into the Node-list through its windows. There are two types of windows, namely Generation Window and Condition Window. Deconverter generates a sentence using the **Word Dictionary**, **Deconversion Rules**, and **Co-occurrence Dictionary** It retrieves relevant dictionary entries from the **Word Dictionary**, operates or inserts nodes by applying **Deconversion Rules**, and makes word selection for natural wording by referring to the **Co-occurrence Dictionary.** The use of the Co-occurrence Dictionary is optional

Deconverter uses the **Condition Windows (CW)** for checking the neighbouring nodes on both sides of the **Generation Windows (GW)** in order to determine whether the neighbouring nodes satisfy the conditions

Sandeep Singh Spall and Parteek Bhatia are with Computer Science & Engineering Department Thapar Institute of Engineering and Technology Patiala, Punjab (e-mail: parteek.bhatia@tiet.ac.in)

for applying a deconversion rule or not. The **Generation Windows (GW)** are used to check two adjacent nodes in order to apply one of the deconversion rules

The word entries of each language are stored in the **Word Dictionary** Each entry of the Word Dictionary is composed of three kinds of elements: the **Headword**, the **Universal Word (UW)** and the **Grammatical Attributes**

A deconversion rule is composed of **Conditions** for the nodes placed on Generation Windows and Condition Windows, and **Actions and/or Operations** for the nodes placed on Generation Windows. The **Co-occurrence Dictionary** provides pragmatic information about the words of a native language First, the deconversion rules are converted into binary format and then binary format rules are loaded The UNL expressions are converted in to semantic net called Node-net. The UWs are replaced with corresponding native language Head Words. If it is not possible to unambiguously decide the correct Head Word for a given UW, Co-occurrence dictionary is used. Co-occurrence dictionary contains more semantic information for proper word selection without the ambiguity. But the use of Co-occurrence dictionary is optional.

Node-net represents the hyper graph (a representation of UNL expressions) that has not yet been visited. Each node contains certain attributes initially loaded from the Language Dictionary and sometime generated by Deconverter during runtime. Each node in the Node-net is traversed and inserted into the Node-list

Node-list shows the current list of nodes that the Deconverter can look at through its windows. Node-list includes two-generation windows circumscribed by condition windows. At the initial stage before any deconversion rule application there are three nodes in the Node-list: Sentence Head node, Entry node and Sentence Tail node. This is explained in Deconverter Specification of UNL Center (UNL Center, 2000). The generation occurs at the generation windows, when the conditions in the condition windows are satisfied The result of rule application is operation on the nodes in Node-list like changing attributes, copy, shift, delete, exchange *etc.* and or insertion of nodes from Node-net to Node-list. The rule application halts when either Left Generation Window reaches the Sentence Tail node or Right Generation Window reached the Sentence Head node. At the end, the nodes in the Node-list represent the generated sentence

III. ARCHITECTURAL DESIGN

There are basically four modules for UNL Punjabi Deconversion: UNL Parser, Case-Marking Module Morphology Generation Module, and Syntax Planning Module. The overall architecture and structure of Punjabi Deconverter has been shown below.

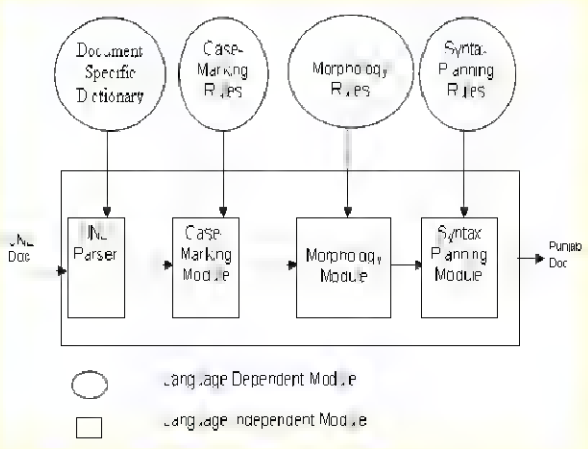


Figure 1.2: Block Diagram of the Deconverter

A. UNL Parser

This module is the important part of any conversion system. This system needs a parser to read an UNL file and convert it into machine understandable format and do some important things.

The parser performs the following tasks:

1. It reports certain errors, if exists, in the input UNL file
2. It instantiates the nodes, scope-nodes and relations present in UNL document.
3. It builds the node-net for every sentence present in UNL document

B. Case-marking Module

Case marker module apply proper case marker for each and every relation in the given UNL expression, *i.e.*, it take into consideration Relational Morphology We follow a rule base approach to incorporate the case markers correctly. A Case Marker data file contains one or more set of constraints for each relation and each of these sets map to different case markers. So, given a node with all its attributes including lexical attributes from dictionary, we search the database for appropriate rule, which the node satisfies and accordingly the case markers are initialized for the case markers

Each line of the Case Marker Database file has 9 columns each separated by a colon, (':') character. Every time a new relation is read from UNL document, this file is referenced once. The 9 fields are described as follows

1 Relation Name
2 Case Marker preceding Parent.
3 Case Marker following Parent
4 Case Marker preceding Child
5 Case Marker following Child.

6. Positive Conditions for Parent
7. Negative Conditions for Parent
8 Positive Conditions for Child
9. Negative Conditions for Child

C. *Morphology Module*

This module is responsible for proper word formation though morphology generation. This module generates most of the words. This module handles noun, verb and adjective morphology generation. This module not only inflects the root words, but also introduces conjunctions, case markers and any other new words if necessary.

UNL relations and attributes govern the morphological rules. Morphological rules due to UNL relations are called relation label morphology

UNL attributes, which express information like aspect, tense; number, gender, speaker's viewpoint *etc.* also play an important role in morphology generation. For example, the attribute @pl means plural. When a noun has an attribute @pl, suffix 'haruu' is added to the stem (noun pronoun). Similarly, if @not is attached to a verb, the verb needs to be negated Suffix 'na' is added at the end of the main predicate verb to negate it.

D. *Syntax Planning Module*

This module is responsible for Punjabi sentence formation by syntax planning. The syntax-planning phase is aimed at generation of proper sequence of words for the Punjabi language. In order to get the correct Punjabi sentence as the output of the Deconverter system, all the rules should be applied in proper order, in other words the proper priority should be given to the rules so that the correct syntax of the Punjabi sentences can be achieved. If there is no proper priority for the rules then it will violate the syntax of the Punjabi sentences.

In UNL relation rel(UW1,UW2), UW1 is the parent node and UW2 is the child node. We plan the syntax, by deciding which child to insert first and at what position (left or right) with respect to other child of its parent This is done by creating a (M+1) x (M+1) priority matrix where M is the total number of relations. We write the relation labels in the first row and first columns. Each *Mij* can be 'L', 'R' or nothing (we represented it by '-'), where *i* is the row number and *j* is the column number

A matrix with dimension 46*46 is made manually Let the Matrix be M. Then any element of the matrix will be denoted by M_{ij} where M_{ij} is element of the i^{th} row and j^{th} column. If the M_{ij} *L* then it means that the position of the child of the i^{th} relation label is left of the child of the j^{th} relation label. Similarly M_{ij} *R* then it means that the position of the child of the i^{th} relation label is right of the child of the j^{th} relation label.

A rank for each relation label is calculated by adding the number of 'R' in the row of each relation label The higher the value of the rank the further right from the main verb is the corresponding word

While making priority matrix for Punjabi, Punjabi translation of the given English sentence and UNL graph of the sentence are observed to decide that when two of any relation will exist then which node with the given relation should be placed left with respect to the other

For example

English: Ram bought an apple for you.

UNL: agt(buy @entry @past,Ram)
obj(buy @entry @past,apple.@def)
ben(buy @entry @past,you)

Punjabi:

ਰਾਮ ਨੇ ਤੇਰੇ ਲਈ ਸੇਬ ਖਰੀਦਿਆ।

In the above example the child of *'obj'*, *'ben'* and *'agt'* relations have the same parent, *i.e.*, 'buy'. Now here as we can see from the Punjabi sentence that 'Ram' comes leftmost, 'you' comes right to 'Ram', 'apple' comes right of 'Ram' and 'you' and buy comes at last. So, when these three relations *'agt'*, *'obj'* and *'ben'* exists, then *'agt'* should be left to the *'ben' and 'obj':* and when *'obj'* and *'ben'* relations exist then *'ben'* relation comes left side

So the priority matrix becomes:

	agt	obj	ben	Rank
agt	-	L	L	0
obj	R	-	R	2
ben	R	L	-	1

Rank for each relation label is calculated by adding the number of 'R' in the row of each relation label. The higher the value of the rank the further right from the main verb is the corresponding word. By using the above priority matrix we can plan the syntax of Punjabi Language from Node-Net, if Node-Net contains *'obj'*, *'agt'*, and *'ben'* relations

IV. CONCLUSION

This paper has described the development of UNL Punjabi Deconverter, a Punjabi language generator Techniques of syntax planning and morphology generation have been used. Syntax planning has been done by studying the syntactic structure of the Punjabi sentences. Morphology has been generated by the effect of UNL relations and attributes on Punjabi word morphology. Most of the information has been generated at morphological level. The current Punjabi Deconverter can deconvert moderately complex UNL expressions. Punjabi Deconverter can be coupled with

other language Enconverter to develop a complete Machine Translation system. It can be used for future UNL Punjabi viewer.

REFERENCES

[1] Dey Kuntal and Bhattacharyya Pushpak, 2003, 'Analysis and generation of Bengali case structure in the Universal Networking Language framework', Proceedings of International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Alexandria, Egypt. http www cfilt iitb ac in convergence03

[2] Dhanabalan T, Geetha T.V., 2003, 'UNL Deconverter for Tamil', Proceedings of International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Alexandria, Egypt http www cfilt iitb ac in convergence03

[3] Uchida Hiroshi, Zhu Meiying, 2002, 'The Universal Networking Language beyond machine translation', Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain. http www-clips imag fr geta User wang-ju.tsai articles/UNL-beyond_MT doc

[4] Uchida Hiroshi, Zhu Meiying, 2005, 'The Universal Networking Language (UNL) specifications 2005', United Nations University, Tokyo http www.undl org unlsys unl/UNL2005

This paper was presented at LRII-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt of India.

8.10 Named Entity Recognition for Telugu

Srikanth. P and Kavi Narayana Murthy, University of Hyderabad, Hyderabad

Abstract— This paper describes our on-going work on Named Entity Recognition (NER) for Telugu. NER involves the identification of named entities such as person names, location names and names of organizations. NER for Indian Languages is a challenging task. There is not much work done in NER for Indian Languages in general and Telugu in particular. Telugu, a language of the Dravidian family, is spoken mainly in southern part of India and ranks second among Indian languages in terms of number of speakers. Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. There are few languages in the world that match Telugu in this regard. NLP Applications such as IR and IE are often hard because of, among other things, the large number of morphological variants for a given root. High performance morphological analyzers have eluded researchers for a long time. In this work we have used news articles in Andhraprabha, a popular Telugu daily. Named entities are generally nouns and it is therefore useful to build a noun identifier. The performance of the Telugu morphological analyzer developed here over the last many years has been evaluated from this angle. Using this in conjunction with a variety of other features we have then built a binary classifier for noun identification using naive Bayes approach. Results are included here. Seed lists of personal suffixes, location suffixes, person name contexts, personal surnames, organization names (such as names of political parties) and place a gazetteer have been developed. A heuristic algorithm for NER has been developed and tested. A training data set of 20,000 words has been developed using the NER system and manual checking. The seed lists have been refined iteratively.

I. INTRODUCTION

NAMED Entity Recognition is a Computational linguistic task in which we seek to classify every word in the corpus as falling into one of four categories: Person, Location, Organization and Not-a-name. In the taxonomy of Computational Linguistics, NER falls under the category of Information Extraction. NER emerged as one of the subtasks of the DARPA-sponsored Message Understanding Conference (MUCs). The task has important significance in the Internet search engines and is an important task in many of the Language Engineering applications such as Machine Translation, Question-Answering systems, Indexing for Information Retrieval and Automatic Summarization.

II. NER IN INDIAN LANGUAGES

There has been a considerable amount of work done for NER in English [1, 2] but these ideas cannot be borrowed straight away for NER in Indian languages. For example, the concept of capitalization does not work for Telugu, whereas it is very useful for English. AU-KBC Research center, Chennai [3] has designed a Biological Name Detection system for English which is based on manually developed rules that rely upon lexical information, linguistic constraints of English and the contextual information. There has not been much work done in NER in Indian Languages. [4] describes a pattern based shallow parsing system of Named Entity Recognition for Bengali. Potential contextual patterns are obtained by taking two contextual words after and before the entity that is already tagged and the patterns that cross a threshold are chosen for further tagging through bootstrapping. There does not seem to be any work on NER in Telugu. Telugu is a free word order Language. Each word in Telugu is inflected for hundreds of word forms. Telugu is primarily a suffixing Language - an inflected word starts with a root and may have several suffixes added to the right Suffixation is not a simple concatenation and morphology of the language is very complex. According to one study (Personal Communication, Uma Maheswar Rao, CALTS, University of Hyderabad) there can be up to 10 lakh forms for a given Telugu verb root. Nouns are also highly inflected. External saMdhi adds to the complexity. This paper is about NER for Telugu.

III. APPROACHES TO NER

Approaches to NER include rule based approaches and Machine Learning techniques. Rule based techniques require substantial linguistic expertise whereas Machine Learning approach requires large amounts of training data. In this work the emphasis is on machine learning techniques.

IV. SENTENCE BOUNDARY DETECTION AND VERB IDENTIFICATION:

Named entities are all generally nouns and recognizing nouns is therefore a useful strategy. We could work by elimination. Telugu is a verb final language and segmenting paragraphs into sentences is therefore useful for recognizing verbs and eliminating them. Full stops and exclamation marks are taken as potential sentence boundary markers and segmentation into sentences effected if the last word has

Srikanth. P and Kavi Narayana Murthy are with the Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, 500046 (e-mail: patilsrik@yahoo.co.in, knmuh@yahoo.com)

4 or more characters. On sample evaluation, 506 out of the 520 sentences detected out of 506 actual sentences given were correct, giving us a precision of 97.3%. Of the 506 sentences, 12 sentences do not end with a verb and we conclude that 97 3% of the sentences end with a verb

V. DEVELOPMENT OF TAGGED DATA

News articles generally start with a headline and the body starts with location name, month and date. A seed list of location names is extracted from this. Seed lists of personal surnames, locational names and organization names have also been developed. Lists of person suffixes such as "reDDi", "na.yuDu" etc, locational suffixes such as "ba d", "pe Ta" etc, and name context lists are maintained for tagging the corpus. It has been observed that whenever a context word (such as "maMtri") appears, then in many cases the following two words (consisting of a surname and person name) indicate a person name. This way we build a list of person names. After extensive experimentation over many iterations, a training data set of 20,000 words has been developed

VI. NOUN IDENTIFICATION

It is useful to recognize nouns and eliminate non-nouns. The Telugu morphological analyzer developed here has been used to obtain the categories. A stop word list including function words has been collected from existing dictionaries and stop words are removed Words with less than three characters are unlikely to be nouns and so eliminated Last word of a sentence is usually a verb and is also eliminated. Digits are eliminated. Verbs are recognized based on a list of verb suffixes and eliminated. Telugu words normally end with a vowel and consonant ending words (laMDan, sTe San, etc.) are usually nouns. Existing dictionaries are also checked for the category. Using these features, a naive Bayes classifier is built using the available tool WEKA. Results are given in the tables 1 and 2

Table 1: Noun Identification using Morphological Analyzer

	NOUN	NOT-NOUN
Precision (%)	95 97	64 59
Recall (%)	64.65	95 96
F-measure (%)	77.25	76 2

VII. Table 2: Noun Identification using a naive Bayes Classifier

	Test-set 1	Test-set 2
Precision (%)	92 91	78.54
Recall (%)	97.26	91.65
F-measure (%)	95 03	84 59

VII. NAMED ENTITY RECOGNITION AND CLASSIFICATION

Nouns detected above are checked against the heuristics described earlier for detection and classification into various types. The system has been tested on a data set of 5280 words Results are given table 3.

Table 3 Name identification using heuristics

	NAMES	PERSON	PLACES
PRECISION(%)	82.22	80.7	84.37
RECALL (%)	70.95	66.2	81.88
FMEASURE(%)	76.15	72.73	83.07

REFERENCES

1. Baluja, S., Mittal, V.O., Sukthankar, R.: Applying machine learning for high performance named-entity extraction Computational Intelligence 16 (2000) 586-596
2. Collins, M., Singer, Y.: Unsupervised models for named entity classication. Michael Collins and Yoram Singer Unsupervised models for named entity classication. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999 (1999)
3. Meenakshi Narayanaswamy, K. E. Ravikumar, V.K.S.: A biological named entity recognizer. Pacic Symposium on Biocomputing, Hawaii (Jan 2003)
4. Eqbal, A.: Named entity recognition for bengali. Satellite Workshop on Language, Artificial Intelligence and Computer Science for Natural Language Applications (LAICS-NLP), Department of Computer Engineering Faculty of Engineering Kasetsart University, Bangkok, Thailand (2006)

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.11 Preparation Problems in Developing Lexical Resources for Computing

Rita Mathur

Abstract— The paper focuses on the various problems that occur inevitably during the process of developing a lexical resource for computing. Since lexical resource is vital for computing, developmental process involves a multifaceted task. It requires at least three kinds of activities pertaining to linguistic, grammatical, and lexicographical areas. Linguistic activity includes morphological and semantic information. Word formation rules and word sense disambiguation are the core of linguistic activity. Tagging for parts of speech requires grammatical activity and selection of canonical forms or lemma is integral part of the lexicographical activity. Thus, development of a lexical resource is a kind of amalgamation of linguistic theory and practice.

Index Terms—Basic form, Morphological form, Tagging

I. INTRODUCTION

MOST of the languages consist of vocabulary or lexical items belonging to different sources. Diachronically they undergo many changes. Numbers of lexical items are borrowed. They become part of the language, the native speakers accept them. Interestingly only basic forms are borrowed, not the inflectional and derivational forms, which means the basic forms loose their morphological forms of the source language and adopt the morphology of the target language, in which they are merged. For instance, Hindi word [jangal] 'forest' is borrowed in English, but the oblique plural [jangalo~] is neither borrowed nor it is a possible form in English. The word has to follow English word formation rules. Similarly English words [rail, doctor] are borrowed into Hindi and are nativized. They inflect for plural and oblique forms as per Hindi morphological rules

In Hindi the vocabulary can be classified into various strata. Initially classification is possible into *native* and *borrowed* words. *Native* words can be defined as inherited by the language from its earlier stages of the development, through phonological change. On the other hand, loan or borrowed words remain same. They retain their form of their source language. These are known as *tadbhava* (originated from) and *tatsama* (same as that) respectively. However, these terms are used for borrowed words from Sanskrit In Hindi these *tatsama* and *tadbhava* words have different morphophonemic patterns.

Another classification may be according to the internal structure, grammatical category, and inflectional and derivational behavior of the lexical items. For instance, in Hindi a noun inflects for gender and number, verb inflects for tense, aspect, and mood and also for gender. Whereas in English verbs do not inflect for gender. Besides inflections, derivation and concatenation of suffixes are to be sorted out

Lexical resources, therefore, are not simply copying or collecting lexical items. It involves the expertise of the compiler who should be well acquainted with linguistic principles and also well versed with the language. There are various kinds of lexical resources, for Instance, word list, machine-readable dictionary, text corpora, thesaurus, word net so on so forth. The list is not exhaustive. This paper would concentrate on the building of a word list and discuss about salient features essential for developing. Nevertheless these are not universal; one has to consider the idiosyncratic featuresof the language that have to be sorted out before taking up such an activity.

II. WORDLIST

Word list may consist of basic forms. Morphological forms may be generated through a programme. This saves space and word list may contain more lexical items. On the other hand, base words as well as morphological words can be stored in the database In the latter case words need not to be generated; the search engine would be able to search and produce the correct output. Nevertheless, in both the cases annotation of the words is very significant. Words are annotated for canonical forms, morphological forms, parts of speech, sense and grammar. Since languages are dynamic and native speakers' pragmatic skills enable them to use words in variety of contexts, lexical items have multiple senses and usages. This is the problematic area if the word list is for machine translation. In that case, mere word list is not sufficient It has to be annotated text corpora, machine-readable dictionary or word net. Care has to be taken for ontology, concordances, etc. Semantic disambiguation and various semantic relations have to be carefully marked

A: ProblemswithMorphology:

There is no universal morphology. Each language has its own closed set of morphological rules. The generalizations are possible only within the language and they do not apply onto various other languages. Thus, in order to develop any lexical resource, morphology of that particular language has to be explored. Morphology of agglutinative language will be quite different from that of an inflectional language Moreover, this will be different for highly inflectional languages. Morphology could be simple and complex.

Rita Mathur is with Alyawar Jung National Institute for the Hearing Handicapped Mumbai (e-mail. ritamathur@yahoo.com)

A complex morphology would be more challenging for computing

Cross-linguistic examinations reveal that there are languages in which grammatical information is presented by the length of a vowel, within the morpheme. In New Zealand Maori (as reported in McCarthy: 1992) Nouns denote plurality by lengthening the vowel of the first syllable For instance,

Singular: tangata 'person' wahine 'woman'

Plural: **taa**ngata 'persons' **waa**hine 'women'

Above examples suggest that word formation rules of the morphology are paired with phonological rules grouped together at various levels. The output of each morphological rule is cycled through the phonological rules of that level. In turn, phonology of that level triggers the next level of morphological rules, paring with phonological rules of that level. In this sense, the rules of morphology and phonology are cyclical. They are made to apply in a cycle, first to the root then out ward to the affixes nearest to the root and then to the outer layer of the affixes. Thus, the "lexical rules" apply within the words whereas "post lexical" rules apply across the boundary.

Hindi Morphological patterns are seen on these distinct levels. They can be defined as *non-concatenative* (level 0, where vowel reduction is seen) and *concatenative* (levels for affixation) of morphology. The influx of phonology is inevitable on all the levels. In classical linguistics this phenomenon is described as 'vowel alternant'. Vowel alternation is a kind of idiosyncratic feature of Hindi. This can be seen in *causativization* of a verb stems This involves the alternation between short and long vowel as well as the concatenation of the affixes. The formation of the verbal sets contains phonological and semantic information Morphologically they are partially similar. Syntactically causative forms reveal direct and indirect causation. Kellog [1887:252] notes this distinction as 'primitive', 'causal' and 'second causal' verbs. Saksena [1982] digresses from the traditional notion, which considers an intransitive verb form to be a base for all the formations and argues that the transitive verb forms, having a long vowel, should be considered as a basic form. The reason being, with intransitive verb form as basic, the morphological generalizations are not possible as the causativization by vowel lengthening is syntactically restricted. For instance, transitive form parh 'to read' might become paarh * which is not possible. Phonologically it is ambiguous to know that short vowel will alternate with long ii , uu or with long /E and O. While via reduction rule long /E and O alters with i and u, respectively. Pray [1976] considers that the majority of verb stems are *'tadbhava'* forms. *'tatsama'* forms are mostly borrowed as nouns or adjectives, not as verbs. Hence, the alternation of tense and lax vowels characterizes many kinds of phonological forms in addition to a sequence, which constitute a stem but not the entire word [Pray: 1976: 96]

The discussion above endorses vowel reduction before concatenation of affixes. It is observed that the causative formation on the Hindi verbs is a systematic process. Following are a few examples:

Level 0: [chil] vt > [chil] vi 'to peel'
Level 1: [[chil]aa]] c [[chil]vaa]] c 'to cause to peel'
L 0 [naac] vi > [nac] 'to dance'
L 1 [[nac]aa]]c > [[nac]vaa]] c to cause to dance'
L 0 [TaTol] > [TaTul] 'to feel'
L 1 [[TaTul]vaa]] c 'to cause to feel'

Another systematic process of non-concatenative morphology is seen in the derivation of nominal forms, diminutive forms, and compounding. For example

lohaa 'Iron' > luhaar 'black smith'
sonaa 'gold' > sunaar 'gold smith'
lotaa 'tumbler' > lutiya 'small tumbler'
khaat 'bed' > khatiya 'small bed'
ghoDa + dOD > ghUD dOD 'horse race'
raajaa + vaaDaa > rajvaaDaa 'royal dwelling'
paanii + chakkii > panchakki 'flour mill, run by hydro energy'

Besides these alternations, morphological and word boundaries should be considered. For instance, in Marathi *khaaNyaasaaThi* 'for eating' is one morphological word, whereas, in Hindi it is a phrase, not a morphological word (*i.e., khaane ke liye*). Another interesting example of word and morpheme boundary can be seen with the pair of words : *premlata* 'a name' and *komalata* 'tenderness'. Though there is partial phonetic similarity but in earlier example there is a word boundary *prem + lata* and in latter case there is a morpheme boundary *komala#ta*. In English, words like *'blackberry', 'cranberry'* and *'raspberry'* put forth the same problem. Thus the concept of word boundary and morpheme boundary must be accounted for. This elaborates the indigenous nature of morphological profile of a language, which is inclusive of morphemic and allomorphic generalizations along with segmentation, productive rules, and typology; if ignored the computing outputs are going to suffer

B. *Problems with Semantics*

Meaning is not a single unitary concept, rather, it is a group of components or complexes. These are perceived as a *gestalt*. It appears in the minds of the speakers with *associative bond*. These semantic components are associated with logically implicational constraints. For Instance, [animate] vs. [non animate], [human] vs. [non human], [male] vs. [female]. At the same time words are related with hyponymy,

homonymy, synonymy, and antonymy relations. Thus one has to consider all the important factors for the selection of lemma

These semantic values are based on ontology, argument structure along with selection restrictions, and semantic relations.

C: Problems with Orthography:

Pronunciation affects the writing. In English, the gap between spoken and writing system is much more in comparison to that of Hindi, Marathi, Spanish, etc. Spelling variation is almost nil in English. But in Hindi there are a lot of possible spelling variations. This makes the task tougher. As the pronunciation is variable, people do not speak in the same way. The knowledge about various dialectal and regional variations becomes a necessity. Moreover, a possible disparity is seen between *'in vogue usage'* and in *'traditional usage'*. This is even found in leading media expressions. Hindi is a potential language for such kind of tiff. The issues involve use of *maatraa, nuktaa, halant, anusvaar, and chandrabindu in* Hindi. Many linguists and language experts opine to keep the variations, since people accept them. On the other hand, purists debate for the correctness. Nevertheless, Hindi orthography is very much affected by modern trends and mixing attitude of the people. Ultimately the compiler has to decide with due considerations to traditional usage, printing and writing conveniences. Secondly, he may be motivated by his own individual preferences. Whatever is selected consistency with pattern remains most important factor.

III. STANDARDISATION

To develop lexical resources methodology, software tools, and standards are to be followed meticulously. Selection of the lexical entry should be in accordance to the needs of computing. There are many standardized formats and software tools that make the task simpler and enable a linguist and or a lexicographer to work on a lexical resource. Nevertheless, compiling of the data should be according to the purpose of the software. Grammatical and orthographical tags in a word list would suffice the needs of spelling checker, parser and so on. But for machine translation semantic disambiguation tags are more important. For that purpose lexical resource must be in the form of annotated text corpora or a word net. Apart from 'purpose' the 'ultimate users' are important. Therefore, word list or any other kind of lexical data must be specialized. A general lexicon may not be sufficient

Keeping purpose and user in mind, generalized rules must be formulated. Again these rules have to have their base in idiosyncratic nature of the language.

Consistency: In order to standardize the word list or any other type of lexical resource, consistency is necessary. Work has been done for the standards of description and creation of the lexical resources, particularly, to facilitate engineering applications (Peters: HTML page) to name a few, as mentioned by Peters, TEI, EAGLES, ISLE, GENELEX, etc. They have defined large numbers of lexical features and tag sets, which can be used to create a meta data for computing Consistency should be observed at all levels. At the level of orthography spelling standardization, regional variation, and style should be selected consistently For Instance, In Hindi, the forms of *khaDI boli* are different from that of other dialects, namely, *purvi* or *bihari*. One must adhere to the variety, which is selected for developing a resource. Standard guidelines are available for the spelling, innovative changes. For instance, in Hindi, care must be taken for problematic issues mentioned above and a consistent pattern according to the standard guidelines should be followed. Personal preferences may be utilized sometimes if the compiler or lexicographer is a native speaker of the language

IV. CONCLUSION

To conclude I would like to emphasize that building of a lexical resource is a joint effort of the programmer and the linguist. Linguist's responsibilities are more because a programmer would bank upon a linguist for the resource. In nutshell following are the area pertaining to various linguistic levels.

- *Phonology* : How words or lexical items are syllabified. In case of stress languages, stress markers should be analyzed.
- *Morphology*: Morpheme and word boundaries must be analyzed Allomorphic rules should be formulated Stem, word formation rules are to be looked for
- *Syntax*: Parts of speech and other syntactic devices viz. inflection and derivation, mass noun vs. count noun, modifiers, their gradable nature, gender, and agreement must be explored
- *Semantics*: Meaning may be analyzed as per lexical meaning, connotational meaning and collocational meaning. Ontology or related concepts must be explored. Homonymy, polysemy, antonymy are to be carefully marked

REFERENCES

[1] Kellog, S H. (1965): A Grammar Of Hindi Language Routledge and Kegan Paul, London

[2] McCarthy,J.J (1981): A Prosodic Theory Of Non-Concatenative Morphology: Linguistic Inquiry , Vol. 12 no.3 373-417

[3] McCarthy, A.C. (1992): Current Morphology Routledge , New York

[4] Pray, B.R. (1970): Topics In Hindi Urdu Grammar, Research Monograph series No. 1, University of California

[5] Saksena,A. (1982): Topics In The Analysis Of Causative With An Account Of Hindi Paradigms University College Press , London

[6] Singh, A B. (1971) On Echo Words In Hindi: Indian Linguistics, Katre Felicitation Vol 30 Deccan College, Poona

[7] Wim Peters: Lexical Resources: HTML page. www.dcs.shef ac uk ~wim BALRIC introduction lang uage resource.pdf

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.12 Corpus-based Statistical Approaches for Stemming Telugu

M. Santhosh Kumar and Kavi Narayana Murthy, University of Hyderabad

Abstract This paper is about Corpus based Statistical approach for Stemming Telugu. Telugu, a language of the Dravidian family, is spoken mainly in southern part of India and ranks second among Indian languages in terms of number of speakers Telugu is a highly inflectional and agglutinating language providing one of the richest and challenging set of linguistic and statistical features. There are few languages in the world that match Telugu in this regard. NLP Applications such as IR and IE are often hard because of, among other things, the large number of morphological variants for any of given term. High performance morphological analyzers have eluded researchers for a long time. This is where stemming comes into picture. Stemming is a technique in which the variant forms of a word are reduced to a common form, thereby also enabling extraction of common suffixes. Thus, words which are different in surface form but have a common stem are conflated. Well known techniques of stemming for other languages including English are rule based or dictionary based They aim at the removal of suffixes based on certain rules or dictionary look-up. For a language like Telugu it is not easy to build a rule based Stemmer because of the large number of morphological variations and unavailability of adequate lexical resources. In this paper we describe several corpus based statistical techniques we have developed for Stemming for Telugu, including n-grams, suffix-trees. We have also developed syllabification rules and have include some syllable statistics here

Index Terms morphology, syllabification, stemmer, n-grams, suffix tree

I. INTRODUCTION

STEMMING is a technique in which the various forms of a word are converted to a common root Stemming is a common language processing task in most information retrieval systems, where it is used to improve the ability to match query and document vocabulary. Both inflectional and derivational morphology lead to multiplication of word forms and stemmers are designed to handle these. We may use a linguistic approach, using prior knowledge of the morphology of the specific language, or a corpus based approach based on statistical principles using a text corpus in the given language. Rules based approaches can be more effective depending on the quality of morphological analysis. However, a linguistic approach implies expert manual labor. Stemming algorithms based on statistical methods require little manual effort.

A number of stemming algorithms have been proposed in the literature. Most of them are rule based or dictionary based It is very difficult to construct a rule based stemmer for a language like Telugu which is so rich in morphology. According to a study by an expert linguist[1] there can be up to 10 lakh different forms for a single Telugu verb. Dictionary based stemmers are also not possible for Telugu since such dictionaries are not available. In recent times, corpus based statistical approaches have emerged as promising alternatives to traditional linguistic approaches. While corpus based and statistical approaches to languages have been well established elsewhere in the world, India is still lagging behind. In this paper we describe several corpus based statistical techniques we have developed for Stemming for Telugu, including n-grams and suffix-trees. We have also developed syllabification rules and have include some syllable statistics here.

[1] Personal Communication, G. Uma Maheshwara Rao, CALTS, University of Hyderabad

M Santosh Kumar and Kavi Narayana Murthy are with the Department of Computer and Information Sciences, University of Hyderabad, Hyderabad Andhra Pradesh (e-mail: santosh_ai@yahoo.co.in, knmuh@yahoo.com)

II. A SURVEY OF STEMMING TECHNIQUES

There are several stemming techniques proposed in the literature Here we survey some of them.

A. Lovins Stemmer

The Lovins Stemmer is a single pass, context-sensitive, longest-match stemmer developed by Julie Beth Lovins of Massachusetts Institute of Technology in 1968. Lovins stemmer maintains a list of most frequent suffixes, 250 in number, and it removes the longest suffix ensuring that the stem is at least 3 characters long. It is quite unreliable and frequently fails to produce the correct stem.

B. Porter Stemmer

The Porter stemmer is a conflation stemmer developed by Martin Porter at the University of Cambridge in 1980 [P1]. Porter stemmer is designed for English language. Porter stemmer is based on the idea that suffixes of words are mostly made up of a combination of smaller and simpler suffixes. It has 5 steps applying rules within each step. If a suffix rule matches a word, then the conditions attached to that rule are tested and the stem is obtained by removing the suffix. Porter stemmer is a linear step stemmer. The Porter Stemmer is readily available and is widely used

C. Dawson's Stemmer

The Dawson Stemmer was developed by J.L. Dawson of the Literary and Linguistics Computing Centre at Cambridge University It is based upon the Lovins stemmer, extending the suffix list to 1200 suffixes. It keeps the longest match and single pass nature of the Lovins stemmer. It replaces recoding rules which were found to be unreliable, using instead, an extension of the partial matching procedure, also defined within the Lovins Paper.

D. Krovertz Stemmer

The Krovertz Stemmer was developed by Bob Krovertz, at the University of Massachusetts, in 1993[P2]. It is based on the morphology. Krovertz stemmer effectively and accurately removes inflectional suffixes and then checks a dictionary.

E. Paice/Husk Stemmer

The Paice-Husk Stemmer was developed by Chris Paice at Lancaster University in the late 1980s and was originally implemented with assistance from Gareth Husk [8]. It is an iterative stemmer. It removes endings from the word in a finite number of steps. It uses a separate file which has different sections for each letter from the alphabet. The sections are ordered alphabetically. An index is built from the last letter of the word. It specifies an ending which matches the last letter of the word. If any special condition for that rule is satisfied (Ex: rule is applicable if no other rules are as yet applied). Application of a rule should not shorten the word by more than specified length

The ideas and techniques used in these stemmers are by and large not applicable to morphologically rich languages such as Telugu. Simple affix-stripping will not suffice as Telugu morphology involves complex morpho-phonemic changes

III. THE LERC-UOH TELUGU CORPUS

A corpus is a large and representative collection of language material stored in a computer processable form. Corpus provides the basic language data from which a variety of lexical resources can be generated. The Telugu corpus developed at the Language Engineering Research Centre (LERC), Department of Computer and Information Sciences, University of Hyderabad, India, referred to as LERC-UoH corpus here, adds up to nearly 39 Million words, perhaps one of the largest corpora for any Indian language today [6]. From this Telugu text corpora, a list of 33,20,920 different word forms with frequencies has been extracted. This list forms the basis of all of our experiments here. It may be observed that available printed dictionaries of Telugu contain 15,000 to 30,000 root words. We thus have some idea about the complex nature of Telugu morphology.

The LERC-UoH corpus is in ISCII encoding. For convenience, we have mapped the word list extracted from this corpus into Roman notation using a tool developed by us here [5].

IV. CORPUS ANALYSIS AND PRE-PROCESSING

A. Basis Of Statistical Stemming

Telugu is primarily a suffixing language - an inflected word starts with a root and may have several suffixes added to the right. Suffixation is not simple concatenation-complex morpho-phonemeic changes occur at the junctures. The main idea in stemming is to divide a given word form into a root and a single suffix-complex including all the suffixes. No attempt is made to analyze the internal structure of the suffix complex

The premise of statistical stemming is that the best place to cut a word into a stem and (a combination of all) suffix (es) is the one that globally maximizes the probability of the root as also that of the (combined) suffix. There are various levels at which this can be done. One could consider words as sequences of symbols, each symbol encoded as a single byte. Byte-level analysis is completely inappropriate for Indian scripts [7]. Orthographic units in Indian scripts are called akSara-s [6] It is really morphemes which combine to form full words according to morpho-phonemic rules of the language and akSara-s are also not appropriate units for morphology or stemming since akSara-s correspond to C*V syllables alone in Telugu. Only syllables - spoken units making up morphemes, can form the right basis for stemming in Telugu. In this work, rules for syllabification have been worked out for Telugu, tested and refined. All of our experiments are based on the assumption that words are sequences of syllables and morpheme boundaries coincide with syllable boundaries.

B. Syllabification

Syllabification is the separation of the words into syllables. Syllables are unit of organization of sequence of speech sounds. It is typically made up of a syllable nucleus (most often a vowel) with optional initial and final margins (typically consonants). Syllables are considered as phonological building blocks of words The written separation is usually marked by hyphen. An example of a Telugu word 'tina:lanukuMTunna·Du' is syllabified and is written as ti-na:-la-nu-kuM-Tun-na.-Du

C. Rules Of Syllabification

The following rules for syllabification in Telugu have been found in literature Here C is a Consonant and V is a Vowel Every syllable is centered at a vowel and surrounding consonants are split as per the following rules: *RULE 1:* Initial and final consonants in a word go with the first and last vowel respectively. *RULE 2:* VCV: The C goes with the right vowel *RULE 3:* 2 or more Cs between Vs: First C goes to the left and the rest to right. Note: Telugu words never show two or more vowels in sequence

We have carried out a small syllabification study with native speakers of Telugu to validate these rules. Based on this study, we propose the following additional rules. Specified consonants followed by y, r, l or v are taken as single compound consonants as given below. Here (x, y) (a) is understood as 'x or 'y' followed

by 'a' *RULE 4:* (Any consonant except y, H, M) (y) is taken as single consonant *RULE 5:* (Any consonant except y, r, l, L, L~, h', n, n~, N, N~, M, H) (r) is taken as a single consonant *RULE 6 :* (k, c, T, t, p, g, j, D, d, b, m, s', S, s) (l) is taken as a single consonant. *RULE 7:* (k, c, T, t, p, g, j, D, d, b, s', S, s, r) (v) is taken as a single consonant. With these rules, the Telugu word list was syllabified to obtain 34,644 distinct syllables.

D. *Coverage Analysis*

Not all syllables occur with equal frequency. A coverage analysis is performed to explore the percentage of words covered by a given number of most frequent syllables. See figure 1. It may be observed that 5000 most frequent syllables account for 96.6% of the words in the whole corpus.

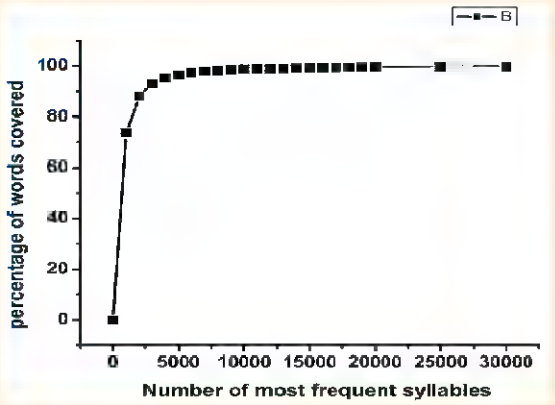


Fig 1: Graph showing the coverage analysis

V. PROPOSED STEMMING TECHNIQUES

A. *Heuristic Stemmer*

We initially explored a variety of heuristics under the premise that *"the best place to cut a word into a root and a suffix is the one that globally maximizes the probability of the root as also that of the suffix".* After extensive experimentation, we have found the following heuristic score to give best results

$$\text{Score} = \frac{(2*P*S)}{(P+S)}$$

P = Frequency of prefix * length of prefix + 0.5.
S = Frequency of suffix * length of suffix + 0.5.
See results later

B. *N-Gram Based Stemmer*

An n-gram is a substring of n consecutive tokens in a stream of tokens. Unigrams are n-grams with n=1, bigrams are n-grams with n=2 and trigrams are n-grams with n=3. Telugu is mainly a suffixing language and hence the initial portions of inflected and derived words generally match with the initial portions of the root word. Exceptions are rare. One exception is: 'ra:' and 'raMDi' are forms of 'vac' Hence we could cluster all the word forms based on the word initial n-grams. Since mono-syllabic words are not many, we have clustered all the word forms based on the word initial bi-grams to obtain 2,67,502 clusters. The smallest word found within a cluster is taken as the root word or lemma

C. *Suffix Tree Approach*

Here words are represented as a suffix tree. For each word and for each possible prefix, the successive *verity* [3] is calculated. Let p be a word of length l and p_i the first i characters of p. Let D be the set of words. $D(p_i)$ is defined as the subset of D containing words whose first i characters match p_i. The successive verity of p_i, denoted by $SV(p_i)$ is thus defined as the number of distinct characters (here syllables) which occupy the $i+1^{th}$ position in the words in $D(p_i)$". See figure 2 At the bottom of the figure *successive verity* for each possible prefix of the word *'tinna:nu'*, are given. Stemming is performed at a position where the *successive variety* is maximum [3] We observe that this criterion does not work well for Telugu. In many cases the maximum value occurs after the very first syllable. After extensive experimentation, we have obtained a set of heuristics to decide which among the first four maxima is to be taken for stemming

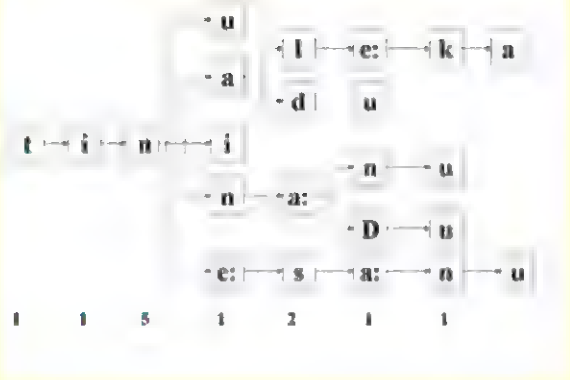


Fig 2 : Suffix tree with successive variety values

VI. PERFORMANCE MEASURES

Usually the performance of a Stemmer is analyzed in terms of its contribution to performance improvement to an IR system. Here we use the following alternative measures of performance.

A. Accuracy

Accuracy is calculated manually over a test data set of 500 words randomly picked from the corpus.

B. Strength Of A Stemmer

(i) *Index Compression Factor (icf)*: icf represents the extent to which a collection of words is reduced by stemming.

$$icf = \frac{N - S}{N}$$

Where N = Number of unique words before stemming and S = Number of unique stems after stemming

(ii) *Number of Words per Conflation Class (wc)* : This is the average number of words reduced to a given stem. This metric is obviously dependent on the number of words processed, but for a word collection of given size, a higher value indicates greater reduction

$$wc = \frac{N}{S}$$

Type of Stemmer	wc	icf	accuracy (%)
Heuristic	5.71	0.68	70.8%
n-gram	12.41	0.69	65.4%
Suffix Tree	2 83	0 51	74 5%

It may be observed that greater reduction does not necessarily imply increased accuracy. In these experiments we find the Suffix Tree approach to be the best in terms of accuracy of stemming

VII. CONCLUSION

In this paper we have proposed several corpus based statistical stemming techniques including heuristic based, n-gram based and suffix tree based techniques for Telugu. All the algorithms have been implemented in Perl under Linux. Experiments and results are included. In all cases, words are viewed as sequences of syllables and we have included syllabification rules which are an improvement over the standard rules found in literature. We plan to explore combinations of these ideas to develop better stemmers. We also plan to use these techniques for developing high quality lexical resources including root words and suffix combinations. Bootstrapping techniques will be explored.

REFERENCES

[1] M.F.Porter, "An algorithm for suffix stripping", Program,14(3),130-137.

[2] Krovertz, "Viewing morphology as an inference process", In proceedings of the 16th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 191-202

[3] Mario A.Nascimento and Adriano C.R da Cunha, "An Experiment Stemming Non-Traditional Text", SPIRE'98, Proceedings,p.75-80.Santa Cruz de La Sierra,Bolivia,Sep 98

[4] Massimo Melucci and Nicola Orio, "A Novel Method for Stemmer Generation Based on Hidden Markov Models", Proceedings of the 12th international conference on Information and Knowledge Management. ACM Press 131-138

[5] http. 202.41.85.117 ~santosh556 cknm.txt

[6] G.Bharadwaja Kumar, Kavi Narayana Murthy, B B Chaudhuri "Statistical Analysis of Telugu Text Corpora", To appear in IJDL., Vol 36, No 2, June 2007.

[7] Kavi Narayana Murthy and G Bharadwaja Kumar, "Language Identification from Small Text Samples", Journal of Quantitative Linguistics, Vol 13, No. 1, 2006 pp. 57-80

[8] Paice C, Husk G., "Another Stemmer", ACM SIGIR Forum 24(3): 566, 1990

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.13 Speech Corpora Development in Indian Languages

Shyamal Kr Das Mandal and Arup Saha
Centre for Development of Advanced Computing (C DAC), Kolkata

***Abstract*—Facts Database + Referential Rules→ Artificial Intelligence (AI). Hence, a collection of standard sample data (or corpora) is crucial for any AI. Similarly, for the development of speech technology and speech research the standard annotated speech database (corpora) plays such an anchoring role. Speech corpora provide the important annotated voice segments to the researcher for making test bed, analysis, and collection of statistics on different speech parameters. Quality Speech Corpora should contain most of the basic elements of speech research like acoustic phonetics and acoustic prosodic; and also the reference data for speech technology development. This paper discusses an approach for creation and management of such Annotated Speech Corpora covering the common needs of speech research and speech technology development including speech recognition, speaker recognition and speech synthesis.**

***Index Terms*—Speech Corpora, Annotation, Informant, Prosody, bi-phones, Suprasegmental.**

I. INTRODUCTION

SPOKEN language interface to computers is a topic that has lured and fascinated engineers and speech scientists for more than five decades. For many, the ability to converse freely with a machine represents the ultimate challenge to Artificial Intelligence, as the production and perception of human speech constitute the highest form of human cognition. In addition to being a challenging topic, spoken language interfaces are fast becoming a necessity. In the current Indian context, the needed skills of both computer literacy as well as conversance with written English in the standard QWERTY English keyboard-oriented interfaces to Personal Computer (PC), restrict the usage of such essential IT access device to only a miniscule fraction of Indian population. Since speech is a primary mode of direct communication amongst human beings, it would be highly desirable from added convenience, as well as empowerment, points of view to facilitate for such direct speech-anchored dialogue between man and machine. Added empowerment by such speech enablement could be the direct local language support in place of English. This demands full-scale research and development in the area of spoken language processing (including speech recognition and speech synthesis) in different Indian Languages. The development of such Speech Research and Technology on Acoustics Phonetics of different spoken Indian Languages demands large speech databases for each The creation of annotated Speech Corpora becomes, therefore, a prime enabler for such research. In case of European and some of the Asian languages the annotated speech corpora and creation methodology are already available in European Lang Resources Assoc [1] and Linguistic Data Consortium (LDC) [2]. In case of English, TIMIT is one of the speech data resources that is used by most of the speech researchers.

While corpora of written text exist in most of the Indian Languages by now, unfortunately none of these exist for their spoken versions. The absence of standard speech corpora in Indian Languages has been one of the major retarding factors slowing down the developments of speech research, particularly related to man-machine-man interactions. In addition, such standard speech corpora can also provide a commonly accessible base line to evaluate the efficacies of current state-of-the-art performance in speech research areas [3]

The major issues for building up such a Speech Corpora are the selection of dialect, content, informants and recording environment [4]. Proper annotation and marking of linguistic and phonetic units are also essential. In the selected dialect, such a speech corpora should adequately cover speaking pattern of native informants, normal as well as under different emotional stresses, for different types of sentences under a variety of discourses

Recently C-DAC, Kolkata has developed basic speech corpora in Bangla, Assamese and Manipuri Language, which is considered only the essential requirement to support for the technology research and development for man-machine-man communication in Indian Languages. The developed Corpora may appear to be adequate only in this limited sense. Primary restriction is towards the manner of speaking (i.e. Text Reading Mode). The database for the study of supra-segmental characteristics may be inadequate for deriving rules for the purpose of generation of fully natural speech It may be extremely difficult to mimic spontaneous human speech mechanically. Spontaneous interactions between speakers are not yet included. In this regards C-DAC Kolkata also published a technical report [5][6] for creating such speech corpora. This paper describes those basic procedures and the required augmentation for possible extension of the corpora to support Spoken Language Processing

II. PROCEDURE

Development of speech corpora involves essentially four steps, namely: (a) Selection of content or Text to be spoken by a number of informants, (b) Recording Speech in Digital Format, (c) Annotation of collected Speech Data and (d) User-friendly management of the recorded data along with the associated text.

Shyamal Kr Das Mandal and Arup Saha are with the Centre for Development of Advanced Computing (C-DAC) Kolkata (e-mail: shyamal dasmandal@kolkatacdac in)

A. *Selection of Content or Text*

There exists adequate experience in text corpora building in most of the major Indian Languages. This is primarily useful for Natural Language Processing (NLP), the backbone of text-oriented technology development, such as Machine Translation (MT), Searching and Information Retrieval, Automatic Summarizations, etc. Unfortunately Spoken Language Processing (SLP has an additional dimension that cannot be addressed purely by mimicking these well-understood NLP procedures; here, one must respect the importance of prosodic information present in spoken language and also the gaps absence of strict grammatical structures of written text. While selecting the contents of Speech Corpora, it is, therefore, essential to adequately address various contrasting differences amongst formalisms of grapheme text vis-a-vis their absence in speech.

However, due to primary emphasis on Speech Corpora contents to various needs of technology development for supporting man-machine-man communication, instead of general linguistic research, the content selection naturally gets governed by those specific technological requirements. In this context, Speech Corpora contents fall largely into following four groups

- Speech Research
- Automatic Speech Recognition
- Speaker Recognition
- Speech Synthesis
- Language Identification

The schematic in Figure1 illustrates how the above diverse needs generally influence the content selection for speech corpora. The rectangular boxes indicate types of targeted contents, whereas, the interconnecting arrowed lines highlight the intricate interactions supporting different technological vis-à-vis research needs

B. *Recording Speech in Digital Format*

The above selected content has to be recorded from different informants in digital domain. The environmental noise plays an important role in recording. On the other hand, the telephone channel voice is bandwidth limited by the channel characteristics. Hence, the recorded samples must contain characteristic noises impregnated within the voice sample. However, to begin with, one can also include normal studio quality recordings

In case of Bangla Speech Corpora, speech is recorded in studio environment directly in digital format using the shure-make dynamic microphone connected to a PC add-on sound card driven by commonly available recording software in 16-bit PCM mono mode with a sampling frequency 22,050 Hz

In case of Bangla corpora, recordings had been made for different age groups of both sexes of Bangla Standard Colloquial Bengali (SCB) speaking informant The selection of the informant had been done based on their test performance of uttering of consonant, vowel, and nasal nominal pair. Adequate web-enabled meta-data of the selected informant is also maintained for later referencing.

C. *Annotation of Recorded Speech Data*

The term annotation covers any descriptive or analytic notations applied over raw speech data. The added annotations may include various kinds of both acoustic and or linguistic information; namely, (a) phoneme, (b) syllable, and (c) word boundaries along with the appropriate (d) Parts of Speech (POS) and or (e) phrase clause markers in the sound spectrum. Usually, there lacks unanimity over the definitions of phoneme-boundaries in any continuous speech However, in the present endeavor, the following phoneme marking conventions have been generally followed

- Consonant phonemes are defined from the beginning of occlusion to the end of Voice-Onset-Time (VOT) for plosives and affricates. Inclusion of VOT here means inclusion of the aperiodic transition, which, inter-alia, is co-articulatory in nature; and hence, may vary for the same consonant in different contexts
- For nasal consonants, trills, laterals and sibilants are defined as the durations of closure or constriction as the case may be.
- Vowel phonemes are defined as the total vocalic region inclusive of the vocalic transitions.
- For Vowel-to-Vowel (VV) transitions the segment boundary is marked at the middle of the transitory part
- For diphthongs the segment includes the prominent target and the long transition.

Semi-automatic tagging software is developed for the acoustic tagging which will give a user-friendly environment to the tager [7].

D. *User-friendly Management of the Recorded Data*

As mentioned earlier, Speech Corpora contents are expected to support four distinct types of investigative research; namely, (a) speech research, (b) automatic speech recognition, (c) speaker recognition, and finally, (d) speech synthesis. Moreover, selected corpora contents of each of the above kinds need to have two interrelated components; namely, text and speech. Such text components are stored as Unicode text and the corresponding speech recordings are stored in PulseCoded Modulation (PCM) in .wav format.

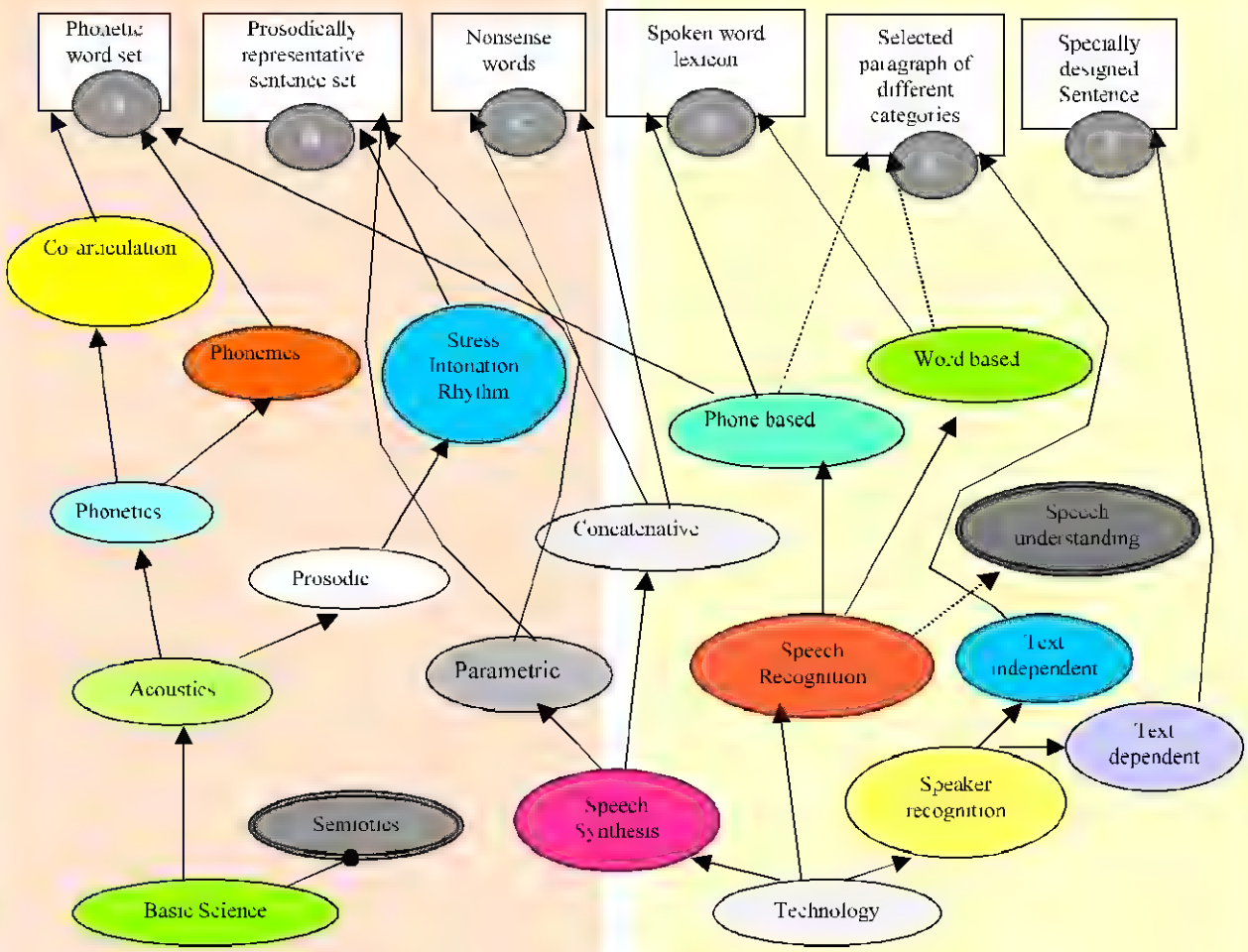


Figure 1 Considerations for Content Selection for Speech Corpora

The tagged notations of the corresponding word or sentences are generated as per the corpora management structure. For convenience, the text and associated speech contents are maintained under the same filename with different file extensions. While the speech recording files are maintained under a separate directory, the corpora management (backed-up with the associated user interface) had been designed in popular XML.

There is a requirement of commonly known phonetic orthographic transcriptions of entire corpora text content to permit viewing under International Phonetic Association (IPA) symbols of the orthographic representations of the same text content.

The corpora should have the two files corresponding to each of the word or sentence for a particular field, one speech sound in .wav and the other tag file with .txt file extension. The name is composed (as in Figure 2, with three component fields; namely, 1) FieldID, 2) Speaker ID, and 3) Word or sentence ID.

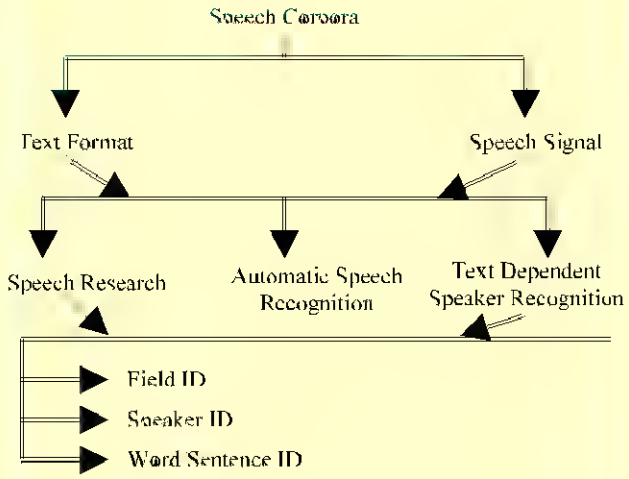


Figure 2 Corpora management tree

III. POSSIBLE EXTENSIONS TO THE SPEECH CORPORA

Developed speech corpus by C-DAC, Kolkata [5] exemplifies the process of building speech corpora for Indian languages. For comprehensiveness, it may be advisable to increase the informants The above corpus is developed for only one of the official dialect for selected languages. However, it may be necessary to build corpora covering the different popular dialects of the languages. There is also a need for special type of corpora to exemplify the speaking style of different speakers. In speech technology development (especially for Speech Recognition), the environmental noise and channel noise also add important attributes, hence, there may be a need that corpora contain sufficient samples of that nature also. The content for acoustic prosodic study in the present corpus is only for one mode of talking The corpus may need to be extended to include other modes (like anger, happiness, joy etc.) Figure 3 describes the possible Extension of the developed basic speech corpora.

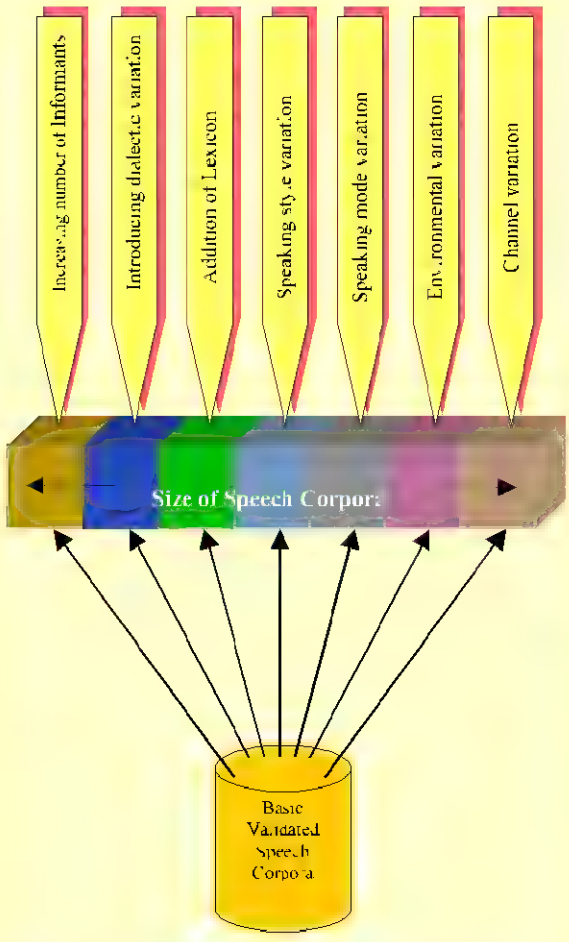


Figure 3: Possible extensions of speech corpora

To develop the channel variation and environmental noise data based for speech recognition technology, existing recorded content is rerecorded through different channel and environmental noise. The developed content is already annotated. So, if we use the same content, then same annotation can be useful after recording through different channel and environmental noise. Only initial time alignment has to be made Following channel variations are incorporated in the existing speech corpora content for speech recognition and speaker recognition technology development.

- Mobile-to-Mobile: The existing speech studio condition recorded content is transmitted through a mobile and after receiving by a mobile it is recorded in computer.
- Mobile to Land Line Phone: in this case the transmitter is Mobile phone and receiver is a Land Line Phone
- Land Line Phone to Mobile: in this case the transmitter is Land Line Phone and receiver is a Mobile phone
- Land Line Phone to Land Line Phone: in this case both the transmitter and receiver are Land Line Phone

All the modulation technology like FDMA, TDMA, and CDMA are covered

In case of environmental noise the above corpora content is recorded from railways station, domestic airport, busy road, and market place.

IV. CONCLUSION

The aforesaid Speech Corpora is considered essential to support the Speech Technological research and development for man-machine-man Interface in Indian Languages. This Corpus may appear to be adequate only within the limited context of its design Primary restriction is towards the manner of speaking (i.e. Standard Text Reading Mode). The database for the study of supra-segmental characteristics may be inadequate for deriving rules for the purpose of generation of fully natural speech It may be extremely difficult to mimic spontaneous human speech mechanically. Spontaneous interactions between speakers are not yet included in the present corpora Furthermore, emotionally rich contents are deliberately avoided In future, speech corpora containing normal spontaneous emotional speech may be designed for both (a) broad based spoken language processing (SLP), as well as (b) linguistic studies on speech Towards this enlarged objectives, such speech corpora may also provide the following

- Normal spontaneous dialogue
- Natural query-answer on various topics of importance in the selected domain.
- Inclusion of dialogues depicting various emotions (data could be collected inter alia from creations of performing arts like Cinema, Television, etc.)
- Inclusion of spontaneous speech (memorable events in the life of informants).

REFERENCES

[1] European Lang Resources Assoc. http www.icp grenet.fr ELRA .

[2] Linguistic Data Consortium. http. www ldc upenn edu

[3] Shyamal Kr. Das Mandal, A.K. Datta, "Annotated Speech Corpora Creation for Bangla", Proceedings of SIMPLE-04, IIT Kargapur, March 2004

[4] Shyamal Kr. Das Mandal, A K. Datta, "Annotated Speech Corpora Creation: An Approach" Proceedings of FRSM-2004, Annamli University, January 2004

[5] Shyamal Kr Das Mandal, Indranil Sarkar, Arup Saha and Asoke Kumar Datta, "Creation of Speech Corpora for Speech Research and Speech Technology Development: An approach", ICSLT-O-COCOSDA-2004, New Delhi, 17-19 Nov 20

[6] Shyamal Das Mandal, Arup Saha, A.K Datta "Annotated Speech Corpora Development in Indian languages" Vishwa Bharat vol. 16, pp49-64, January 2005

[7] Shyamal Kr Das Mandal, Arup Saha and Asoke Kumar Datta, "A semi-automatic speech signal annotation system" Proceedings of FRSM-2006, January 9, 2006, pp. 157-163

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.14 Automatic Construction of Telugu Thesaurus from Available Lexical Resources

M. Santhosh Kumar and Kavi Narayana Murthy, University of Hyderabad

Abstract This paper is about automatic construction of a Telugu Thesaurus from a variety of available lexical resources. A thesaurus links semantically related words and helps in the selection of most appropriate words for given contexts. A thesaurus contains synonyms (words which have basically the same meaning) and as such, an important tool for many applications in NLP too. Yet many of the major languages of India have no thesauri till date. Recent work has focused on automatic or semi-automatic construction of thesauri from annotated corpora and other available lexical resources. Annotated corpora and other lexical resources are limited in Indian languages, and hence, many of the techniques used for other languages of the world are not applicable at present to Indian languages. However bilingual dictionaries are available It has been shown that a thesaurus can be constructed automatically and efficiently from bilingual dictionaries with little human effort [11]. Here we present the algorithm used for this and show its application for the automatic construction of a Telugu thesaurus from available lexical resources. We include examples from the Telugu thesaurus so constructed

Index Terms Thesaurus, Dictionary, Indexing

I. INTRODUCTION

IN very general terms, a thesaurus has been defined as a treasury or a storehouse; hence, a repository, especially of knowledge; often applied to a comprehensive work, like a dictionary or encyclopedia More, specifically, a thesaurus is a book containing a classified list of synonyms, organized to help you find the word you want but cannot think of

We go to thesaurus when we have an idea, some concept or meaning in our mind but we are unable to get just the right word that fits our need. We have some word on hand but we some how feel that there should be a better word, a word that says more precisely what we wish to say, a word that is best for the current context. A thesaurus usually contains an index from where we can start. We look up the index for the tentative word we have with us, a word that approximates what we wish to say but not quite exactly. The index tells us which location in the thesaurus we need to look up. We go to those locations and hopefully we will get the word that we are looking for. At times, we get more ideas and we may want to continue searching from the words we got and we may go on several rounds in this fashion. Given this broad idea, it is not necessary that a thesaurus be constructed strictly in terms of synonyms. Any word that is semantically related in some way to the given word can be linked. In fact by going beyond the strict notion of synonym, we may be able to produce a more general and more useful resource.

II. AUTOMATIC CONSTRUCTION OF THESAURUS

The biggest challenge in constructing a thesaurus, therefore, is in identifying the words that are semantically related to one another. Manual construction of thesauri is a tedious and time consuming task. Manually constructed thesauri also tend to suffer from problems of bias, inconsistency and limited coverage. In addition, thesaurus developers can not keep up with constantly evolving language usage and cannot afford to build new thesauri for many new sub-domains that NLP techniques are being applied to. There is a clear need for automatic construction of thesauri

Recent work has focused on automatic or semi-automatic construction of thesauri from parallel corpora, annotated corpora, and other available resources. See for example [4, 2, 1, 3, 10, 5]. However, these techniques are not applicable to Indian languages at present since corpora and other lexical resources available in electronic form are extremely limited, although there is some recent interest in developing such resources. There are small (about 3 Million words) text corpora for most major languages of India but hardly any parallel corpora or annotated corpora. There are of course no word nets etc. as yet. There are no significant computational grammars or syntactic parsers for any of these languages. Electronic dictionaries are, however, available in many languages

Here we show that a bilingual dictionary is one good source that can be tapped. Dictionaries are more readily available in Indian languages compared to other forms of electronic resources. A bilingual dictionary, especially of the kind developed with applications such as automatic translation, tends to list target language equivalents for each source language word. In doing so, these dictionaries actually group together related words It should therefore be possible to extract this hidden structure and build a thesaurus. This is the main idea in this paper

III. DICTIONARY VERSUS THESAURUS

The content of a thesaurus is very similar to that of a dictionary. A dictionary is typically organized in, say, alphabetical order so that you can quickly locate the word of interest and then you can get the correct

M Santosh Kumar and Kavi Narayana Murthy are with the Department of Computer and Information Sciences University of Hyderabad, Hyderabad Andhra Pradesh (e-mail: santosh at@yahoo co in, knmuh@yahoo com)

spelling, pronunciation, meanings, usage, etymology and other such pieces of information associated with the word in question [P6,P7] A thesaurus, on the other hand, could be organized in terms of an ontology - a hierarchy of concepts, and the words are structured into groups that convey a specific meaning. The difference between a dictionary and a thesaurus, therefore, is more of structure and organization rather than that of content. Both the dictionary and the thesaurus contain words of a given language and their meanings

Given this, it makes a lot of sense to consider a dictionary and a thesaurus as simply two different views of the same data, rather than as two entirely different entities. It appears to be a good idea to store the words only once and provide two different indexing mechanisms, one to use the words as a dictionary, and another to use the same words as a thesaurus [P9] Some kind of a thesaurus can thus be automatically and very efficiently constructed from a dictionary and such a thesaurus can be practically very useful. In this paper we show that a thesaurus can be constructed automatically and efficiently from a bilingual dictionary with little human labor. We show examples from a Telugu thesaurus constructed automatically from bilingual English-Telugu dictionaries. It may be noted that there is hardly any large scale lexical resource available today for Telugu, although Telugu is a major language spoken by many people in India. An automatically constructed thesaurus may not be as good as one that is carefully handcrafted by lexicographers. But it can serve an immediate need. Also, a thesaurus so generated can be viewed as a raw material for further research and development.

IV. METHODOLOGY

To construct a thesaurus automatically, the data needed include the words of the language, the grammatical categories and other relevant features, and the meanings. Different words may have same spellings and a word may have many meanings (homonymy and polysemy). It is important to keep these things in mind while developing a thesaurus. Perhaps the best single source of all these required pieces of information is the dictionary itself. We now give the skeleton of an algorithm to show the basic idea

#ALGORITHM

#INPUT: A DICTIONARY
#OUTPUT: A THESAURUS

#First Create a Reverse Index.

For each dictionary entry with head word W
  For each category i C1, C2, .. Cn
    For each meaning j M1, M2, ... Mp
      For each synonym k S1, S2, ... Sq
        push W into index(i,j,k)

#Create the thesaurus index

For each word W
For all HW index(i,j,W)
synset(W,i,j) synset(W,i,j) Union (i,j,X) for all index(i,j,X) = HW

Note that the algorithm keeps the synsets separately for each category and each meaning, and thus, users should be able to locate the word they are looking for without mixing up different grammatical categories or different senses of a given word

The algorithm has been implemented efficiently using suitable data structures and hashing techniques

V. THESAURUS FOR TELUGU

Telugu is the second most spoken language in India, one of the twenty-two official languages of the Republic of India and one of the official languages of the state of Andhra Pradesh. Telugu has a vast and rich literature dating back to many centuries. Yet there is no widely available electronic thesaurus till date. In this work, a thesaurus for Telugu was generated automatically starting from two English-Telugu dictionaries. One was of these was developed by C.P Brown and the other was developed here as a part of a machine aided translation project earlier. These dictionaries give more or less substitutable equivalents rather than elaborate descriptions or precise definitions and are therefore suitable for our purpose here

More than 30,000 root words were extracted from the above two bilingual dictionaries. Inverted indexing as described above resulted in the thesaurus in less than a second. Below we show excerpts from the thesaurus. For each word in Telugu, corresponding synonyms are listed based on their category and also its sense in English

Synset	Category	Sense
telivi	N	Judiciousness
buddhi sukSmata naipuNi	N	Sagacity
sami:kSanamu ja.gratta	N	Discretion
Kaus`alamu pravi:Nata ne.rpu	N	Skill
jN~a:namu buddhi telivi	N	wisdom

Table 1: Synonyms of the word "vive: kamu" (wisdom).

Here are some statistics. Total number of Telugu words is 30361. Average number of synonyms per word is 1.39. This could be higher if the dictionaries gave more number of equivalents. Maximum number of synonyms for a word in the thesaurus is 28 (for the word 'peMcu', see table-2). Maximum categories for a word are 5 (for the word 'le:du' - Adjective, Noun, Adverb, Determiner and Interjection, see table-3). The synset with maximum number of synonyms for a word in particular category is 9 (see table-4). The total number of synsets found in the thesaurus is 27558.

Table 2: synset of the word "peMcu", is the biggest synset among all the synset in the developed thesaurus

Synset	Category	Sense
sa·gu_ce:yu	VT	cultivate
perugu edugu	VT	Grow
ekkuva ceyyi vRddhi ceyyi vRddhı poMdıMcu	V	Increase
vRddhi ceyyi pogaDu ki:rtiMcu vRdhdhi ceyyı peMcı_cu:piMcu	V	magnify
po:SiMcu	V	nourish
po:SiMcu pro:tsahiMcu	V	foster
po:SiMcu	V	cherish
hecca veyyi guNıMcu abhıvRdhdı_avu	V	multiply
sa gadi yu	V	elongate
merugu_parucu	V	enhances
paikettu pogaDu	V	exalt
po:SiMcu vRddhi ceyyi pado nnati kalpıMcu protsahıMcu	V	promote
merugu_parucu	V	enhance
balaMga:_toyyi	V	boost
po:SıMcu ne:rpıMcu	V	nurture
hecciMcu vRddhi ce:yu nilabeTTu le:vanettu	V	raise

Table 3: Synonyms of the word "le du (no)" which is accounting for 5 categories

Synset	Category	Sense
ka·du	Det	No
ka:du ka:du	Adv	no Not
ka:du paiga le:kuMDaTaM	N	No Nay Haven`t
ka:du	I	nope
e di_ka:du	Adj	no

Table 4 A synset which contains maximum number of synonyms for a category

Synset	Category	Sense
muTTaDi avaro·dhaM aDDaMkı digbadhaM a:TaMkaM muTTaDiMcu dıgbaMdhaM ceyyi dıgbaMdhaM_ce:yu cuTTu_muTTaDi	N	blockade

VI. CONCLUSION

It has been shown earlier that a thesaurus can be automatically and efficiently constructed from a good dictionary with little human effort. In this paper we show the application of this idea for the automatic construction of a Telugu thesaurus. The necessary code was developed in Perl under Linux. The program takes less than a second to construct the thesaurus. The method holds promise since it is relatively easy to develop electronic dictionaries and other lexical resources not yet available for many Indian languages. The quality of the thesaurus depends on the quality of the dictionary we start from. It is also possible to use this tool to verify the quality of a dictionary and hence correct, enhance, enrich and otherwise improve the dictionary itself. The automatically constructed thesaurus can also be taken as a starting point for developing a better thesaurus.

We have shown extracts from the Telugu thesaurus constructed automatically from existing English-Telugu dictionaries. To the best of our knowledge, there was no thesaurus for the Telugu so far. Only informal and limited manual evaluations have been carried out so far but the results are very encouraging. Lack of other thesauri, word-nets, sense tagged corpora, parallel corpora etc. for Telugu is a serious issue for large scale quantitative evaluation of the current work. Lexical resources for Telugu are slowly getting developed and systematic, large scale quantitative evaluations may be possible in future

REFERENCES

[1] W. Z.Chen, S.Liu, L.WenYin, G.Pu, and W.Y Ma. "Building a web thesaurus from web link structure" Technical Report MSR-TR-2003-10, Microsoft Research, 2003.

[2] J.Curran "Ensemble methods for automatic thesaurus extraction" In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing Philadelphia, pages 222-229, PA, USA, 2002.

[3] H.Dejean, E.Gaussier, and F.Sadat. "Bilingual terminology extraction: an approach on a multilingual thesaurus applicable to comparable corpora" In Proceedings of COLING-2002, 2002.

[4] E.A Fox, J.T.Nutter, T.Ahlswede, M Evens, and J Markowitz. "Building a large thesaurus for information retrieval" In Proceedings of the Second Conference on Applied Natural Language Processing, pages 101-108, Austin, TX, 1988 ACL.

[5] J Jannink and G.Wiederhold. "Thesaurus entry extraction from an on-line dictionary" 1999

[6] Narayana Murthy Kavi "Electronic dictionaries and computational tools" Linguistics Today, 1(1) 34-50, 1997.

[7] Narayana Murthy Kavi. "An indexing technique for efficient retrieval from large dictionaries" Proceedings of National Conference on Information Technology NCIT-97, 21-23 December 1997, Bhubaneswar, 1997.

[8] Narayana Murthy Kavi. "Mat: A machine assisted translation system" Proceedings of the Fifth Natural Language Pacific Rim Symposium, NLPRS-99, 5-7 November, Beijing, China, 1999

[9] Sivasankara Reddy A, Narayana Murthy Kavi and Vasudev Varma "Object oriented multipurpose lexicon" International Journal of Communication, 6(1 and 2):69-84, 1996.

[10] T.Takenobu, I.Makoto, and T.Hozumi. "Automatic thesaurus construction based on grammatical relations" In Proceedings of IJCAI-95, 1995.

[11] Kavi Narayana Murthy, "On Automatic Construction of a Thesaurus" Proceedings of ICSLT-O-COCOSDA 2004 International Conference, Vol-1, pp 191-194, November 2004, New Delhi

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.15 The Structure of a Dialect Dictionary of Agricultural Vocabulary in Tamil

S Raja, Annamalai University, Annamalai Nagar, Tamilnadu, India

Abstract **In the present scenario of science and technology advancements several indigenous occupational vocabularies are being replaced gradually by English. This often leads to replacement of culturally related vocabulary like occupational items with textural hybridity. In this context, this paper mainly focuses on various issues related to textural hybridity in Tamil agricultural vocabulary. Apart from this, this paper shows the significance of dialect surveys and the compilation of a dialect dictionary. The significance of dialect surveys is to bring out the variant forms of different regions. Recent research in the field of metalexicography, based on the structure of the dictionaries identified by Hausmann and Wiegand (1989), and Weigand (1989), are made use of in this study.**

I. INTRODUCTION

This paper shows the significance of dialect surveys and the compilation of a dialect dictionary. The significance of dialect surveys is to bring out the variant forms of different regions. Recent research in the field of metalexicography has focused on the structure of the dictionaries identified by Hausmann and Wiegand (1989) and Weigand (1989)

The proposed dictionary will examine the different structures of a dictionary the macrostructure. Here the macrostructure is the selection of lemma sign in the spoken variety only. The microstructure of the proposed dictionary has two important structures; the information structure is not a lexicographic definition but the lexical description given by the farmers in Tamil Another important structure in the microstructure is providing the variation found in the different dialect regions. This is the main aim of the proposed dictionary. For this, the microstructure may be different when compared with the microstructure of a general monolingual dictionary. Here all the variant forms are clearly marked by different labels. These variants are treated in the microstructure as dialect variant form (DVF).

Lexicography is an applied field, the theoretical background for which is provided by linguistics. Almost all the branches of linguistics provide information for the making of a dictionary. Phonetics and phonology are useful for providing information about the pronunciation standardization of the script and alterations in the form of headwords. Grammar (morphology and syntax) is essential for providing grammatical information of different types, like parts of speech of words, gender distinction and causal relation of nouns, conjugation of verbs, etc. Dialectology, Sociolinguistics, Psycholinguistics etc. are useful in giving usage and other types of labels to the headwords. But the branch of linguistics which is most important for the compilation of dictionaries is semantics, especially lexical semantics, as meaning is the central and vital component in the structure of a dictionary entry. This paper mainly focuses on the method of preparation of an article or dictionary entry for agricultural vocabulary in Tamil

II. DICTIONARY DEFINITION

According to the Oxford English Dictionary (OED), the word 'dictionary' is derived from the Latin form 'dictionarius', used in 1225 by English poet and grammarian Joannes de Garlandia (John of Garland) as the title of a collection of Latin vocables' and the word 'dictionarium' is used about a century later by Peter Bechorius (or Becharius). The first recorded appearance of the word dictionary as such is dated to 1526 by the OED; later the word was used by Thomas Elyot in 1538. The French word dictionnaire seems to have been used for the first time by Robert Estienne in 1539 (Bejoint, 2004).

In Webster's third new international dictionary (W3) the definition of dictionary is slightly more detailed, but very similar: 'A reference book containing words usually alphabetically arranged along with information about their forms, pronunciation, functions, etymologies, meanings, and syntactic and idiomatic use '

Matore (1968) finds all definitions of dictionary in dictionaries less than satisfactory. The only one that satisfies him is the one that is proposed by the OED. It runs: 'A book dealing with the individual words of a language (or certain specified classes of them), so as to set forth their orthography, pronunciation, signification and use, their synonyms, derivation and history, or at least some of these facts; for convenience of reference, the words are arranged in some stated order, now, in most languages, alphabetical; and in larger dictionaries the information given is illustrated by questions from literature'

Zgusta (1971) uses a slightly more precise formulation: 'A dictionary is a systematically arranged list of socialized linguistic forms compiled from the speech-habits of a given speech-community and commented on by the author in such a way that a qualified reader understands the meaning of each separate form, and is informed of the relevant facts concerning the function of that form in its community'.

S. Raja is with CAS in Linguistics, Annamalai University Annamalai Nagar Tamilnadu India (e-mail. kiranraja_lng@yahoo.co.in)

III. DICTIONARY TYPOLOGY

Zgusta (1971) opined that when a lexicographer sets out to compile a dictionary, he has to take two basic decisions(1) what part of the total vocabulary of a language the proposed dictionary will cover and (2) of what type the proposed dictionary will belong Both the aspects of typological classification of dictionaries are quite useful in understanding the classification and locating the proposed dictionary. Dictionaries can be classified into various types on the basis of different criteria. Zgusta classifies dictionaries into two major divisions namely, linguistic and non-linguistic dictionaries. The linguistic dictionaries are concerned with the words or lexical units of languages, called word books. The non-linguistic dictionaries are not concerned with the words, but with realia or denotata (thing) and they are called encyclopedias, or thing books.

Classification of linguistic dictionaries has been attempted by a number of scholars such as Shcherba (1940), Sebeok (1962), Malkiel (1967), Comyn (1967), Gelb (1968), Zgusta (1971), Al Kasimi (1977), Singh (1982), and Svensen (1993). Contrasting features of dictionary typology are divided into two basic classes, viz internal features and external features The internal features are concerned with the aspects of the nature of language such as paradigmatic vs syntagmatic contrast, aspects of form meaning, time and area of vocabulary covered. External features are concerned with the target group or user of the dictionary for whom the dictionary is compiled *i.e.*, size, purpose, arrangement of entries or words, number of languages, etc. Internal features are theoretical in nature, whereas the external features are applicational or concerned with the use or practical utility of the dictionary.

IV. METHODOLOGY

Agricultural vocabulary have been collected from fifty points comprising of the five major dialect regions of Tamil Nadu, namely, (1) Eastern dialect, (2) Western dialect, (3) Northern dialect, (4) Southern dialect, and (5) Central dialect. The distance between each point was twenty five to thirty kilometres distance. Informants were selected from densely populated areas. The informants are of different age groups and caste groups, both male and female. Old persons in the age group of fifty to seventy who have no educational background were also selected Lexical items relating to various agricultural areas like soil, irrigation, water lifting instruments, weather and season, ploughing, leveling the field, seeds, crops, harvesting, manure, cattle, etc were included in the detailed questionnaire which was prepared for this purpose. The questionnaire contained 527 direct questions. In addition to this, a number of additional questions were also asked to collect sufficient data. The data were recorded in a tape-recorder. Data were analyzed and presented in the form of a dictionary format.

V. DESCRIPTION OF LEXICAL MEANING

Zgusta (1971) says that there are four basic instruments which are used for the description of lexical meanings. They are: (1) The lexicographic definition, (2) synonyms, (3) exemplification, and (4) glosses and labels. A number of other theoreticians include synonyms as a kind of definition (Svensen, 1993, Kipfer, 1984 Landau, 1989; Hartmann and James, 1998) What Zgusta calls exemplication, others such as Landau (1989) refer to as illustrative quotations, Ilson (1986) as illustrative phrase, Svensen (1993) as examples of usage and Kipfer (1984) calls illustration or exemplification used as the organic part of the definition within parenthesis.

Label is another instrument used to describe the lexical meaning, especially in the compilation of a dialect dictionary (proposed dictionary) and in connotation. Zgusta regards labels as a species of glosses. But others such as Svensen (1993) discuss it separately, while Landau (1989) and Kipfer (1984) describe it as a part of usage'

VI. LEXICOGRAPHIC DEFINITION

There are different types of lexicographic definitions (1) Paraphrase, (2) Referential analytical intentional definition, (3) extensional definition, (4) formulaic definition, and (5) definition by synonyms. The second type, what Zgusta (1971) calls lexicographic definition, Ilson (1986) calls as referential or analytical definition, while Svensen (1993) calls intentional definition, which are modified versions of lexical definitions 'genus proximum' and 'differentia specifica'. The concept represented by the headword is called the definiendum and the definition (verbal description) is called the definiens. Intention denotes the content of the concept which is defined as the combination of distinctive features or what Zgusta calls criterial features which the concept comprises. It expresses a generic conceptual relationship whereby concepts are arranged in classes according to similarities and differences noted between them This has resulted in a hierarchical system with super-ordinate (hyperonym), sub-ordinate (hyponym), and co-ordinate (co-hyponym, concepts. The process of definition involves stating the super ordinate concept next to definiendum, i.e., (genus proximum) together with at least one distinctive feature typical of the definiendum (differentia specificum) (Svensen, 1993) This type of

definition is useful to explain a class of related words constituting a set of hyponymy with or without superordinates, meronyms and also preparing a dialect dictionary

e.g., *kiLai* (n) 'marattin oru paakam1 pakuti'

'branch is a part of tree'

Ital (n) 'puuvin oru pakutil paakam'

'petal is a part of flower'

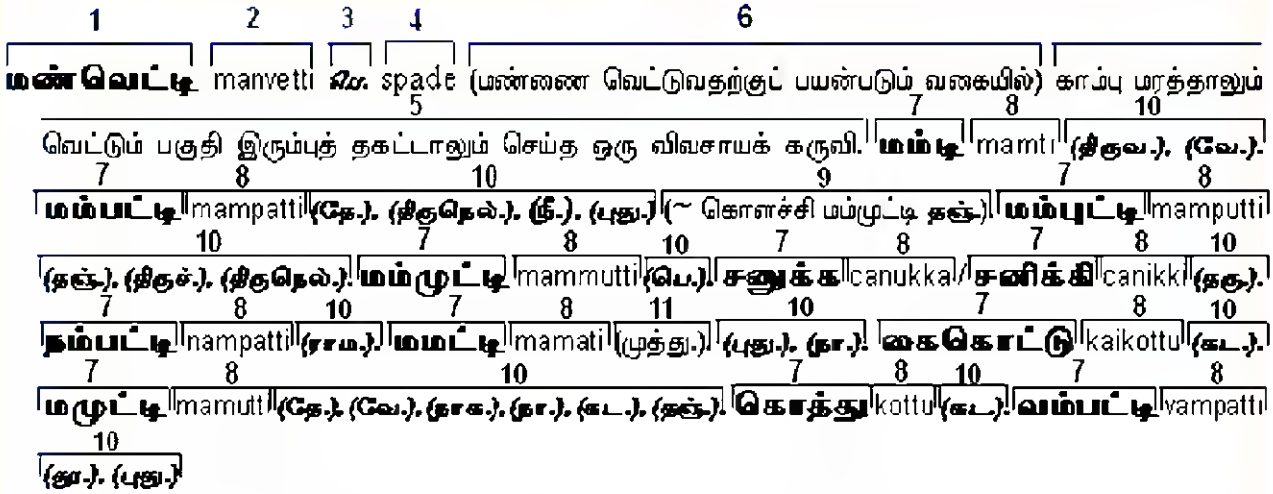
VII. COMPONENT PARTS AND STRUCTURE OF A DICTIONARY

The present article focuses on the dictionary type in general and very particular to the structure of an entry The word 'dictionary' has two meanings in a textural manner, i.e., (1) the whole book and (2) the word list, which constitutes the main part of the book. Both of them have two structures, the textural book structure and the textural word list structure, respectively. The word has several important units and structures. The basic unit of the dictionary is the treatment unit The treatment units have a form and information relating to that form are brought together. The relation of form and information is that of topic and comment (Hausmann and Wiegand, 1989).

A form and information relating to a form are brought together under the addressing procedure. Each information item is addressed to a form called address. In any dictionary the most important item is the definition, but there may be hundreds of other information types, i.e., items. The most important address is lemma (headword or entry word), because the lemma belongs to the alphabetical access structure of the dictionary. Normally all the ordered set of lemmata of the dictionary forms the macrostructure. The lemma and the whole set of information items, which are addressed to the lemma, form the dictionary article The macrostructure is the relative arrangement of the stock of lemmata in the word list (cf. Svensen 1993, Bengenholtz 1995). The macrostructure is the overall list structure which allows the complier and the user to locate information in a reference work. "It is the macrostructure that determines under which lemma the lexicographical item is to be found" (Hausmann & Wiegand 1986). The macrostructure may have a single central word list or additional word lists. The arrangement of words in the macrostructure can be fully alphabetic or systematic (conceptual grouping) or a combination of both. The macrostructure of the proposed dictionary has a central word list. Normally speaking, the structure of information within the article is called the microstructure

The proposed dictionary entry is given in the following format. The headword or article is given in Tamil script with bold letters, followed by the phonemic script (transliteration). The grammatical indication is also given and in addition to this, the English equivalent is also provided. After that, the lexicographic description is given in Tamil, which is added for better understanding of the target language group (Tamil language group). Many types of dictionaries will have the same structure but the proposed dialect has a different structure. That is, in addition to the lexicographic description, the variations found in the different regions are also presented in the same entry. Those variations are labeled. Such type of arrangement is called niching. It means a strict-alphabetical clustering of lemmata or articles that may or may not be semantically related. All the variant forms also find a place in the dictionary entry. There the description or definitions are cross-referenced. If there is no variation in any district to the headword, then it is believed that the speakers of particular district use the same form (i.e., the central region form). In such cases, no form is given in that entry. Since the central region (Thanjavur) has the standard spoken variety, the headword is given in that variety only. All the regions are labeled according to the nesting method. The sample entries are also given.

VIII. GUIDE TO THE INNER STRUCTURE OF THE PROPOSED DICTIONARY



மொளக்குச்சி அடி molakkucci aṭi பெ. புதிதாக நடும் தேயிலை செடி சாயாமல் இணைத்துக் கட்டுவதற்காக செடி ஒரம் நடப்படும் சிறுகுச்சி. (நீ.)

1. Head word entry word (in BOLD script)
2. Phonemic Transcription
3. Grammatical Item
4. English Equivalent
5. Definition
6. Semantic gloss
7. Lexical variation
8. Phonemic transcription of variant form
9. ~ cross-reference
10. Label Item
11. Sub-label Item

IX. CONCLUSION

Due to the influence of science and technology, many new technical terms in English come into existence in almost all fields of Tamil. Because of this reason, much native vocabulary of indigenous occupations are slowly getting replaced by English language. This may lead to permanent loss of the rich occupational vocabulary in Tamil. This study will be quite helpful in preserving the occupational vocabulary in Tamil and recording it in a dictionary format for future purpose. This study may provide a model for the future survey and to record the various occupational vocabularies in Tamil

APPENDIX

SAMPLE ENTRIES OF THE TAMIL AGRICULTURAL DICTIONARY

கழிகோர மடமைடிசட *பெ.* கோரை அறுவடை செய்யும் போது பயனற்றுப் போனக் கோரை. (நா.)

கழிச்சிவிடு மபடஅஉஎனை"ர *வி.* (பார்க்க–அண்டகழி). (வே)

கழிச்சிவுடு மபடஅஉஎர"ர *வி.* சுத்தம் செய்யும் பொருட்டு மண் வெட்டியால் வரப்பை கழித்து வயலின் உள்பக்கம் போடுதல். (பெ.)

கழிமட மபடூஅப"ப *பெ* (பார்க்க–வடிமட). (நா)

கழுத்துக்கட்டி மபடரவவரமஅப""ை *பெ.* மாட்டை ஒன்றோடு ஒன்று/ஒன்றில் பிணைத்துக் கட்டுவதற்காக மாட்டின் கழுத்தில் கட்டப்படும் கயிறு. **தலகயிறு** வபடபமஅலசைர (நா.). **தலகயிறு** வபடபமஅலசைர (பெ)

கள மபடய *பெ.* நெற்பயிரில் முளைக்கும் பயிர் அல்லாத செடி. (திருவ.), (பெ), (தஞ்)

களஎடு மடடயந'ர *வி.* 1) நெற்பயிரில் முளைத்துள்ள பயிர் அல்லாத செடி புற்களை வேரோடு எடுத்தல். (பார்க்க–களபறி). (தரு.), (தஞ்), (கூட.), (ம.), (விரு.), (தே.), (பெ), (சிவ). **களபறி** மபடப'யசை.**களஅளசு** மடடயபடடஉர (தரு). **களபறி** மபடப'பசை (தஞ்.)., (திருவ.), (வே.). **களபெறுக்கு** மயடய'நசுரமமர,**களவாங்கு** மபடபஎடபசூமர (வே.). **களதடவு** மயடயஉய"டஎர, **களதொலவு** மபடபவடிடயர (நா.). 2) சேற்று வயலில் ஏர் உழுதப்பின் அழுகி கிடக்கும் புல் மற்றும் தூசுகளை கைகளால் அறித்தல். (தூ.). 3) களக்கொத்தியால் கொத்தி களைகளை எடுத்தல். (நீ.). 4) மரவள்ளிக் கிழங்கு நன்றாகவிட்டு வளர்வதற்காக அதன் வேர்ப் பகுதியை களைக்கொட்டினால் கொத்தி விடுதல். (நா.)

களஎடுக்கும்மிஷின் மபடபந"ரமமரஅஅ ை *பெ* கயறுகட்டி வரிசை வரிசையாக நடும் புதுமுறை ஒற்றை நாற்று நடவில் முளைத்துள்ள களைகளை எடுக்கும் இயந்திரம்.

களக்கநடு மயடபமமயமே"ர *வி.* இடைவெளிவிட்டு அகலமாக நடுதல். (திருநெல்)

களக்கொள்ளி மபடபமமடிடட *பெ.* களைகளை அழிப்பதற்காகப் பயன்படும் ஒருவகை இரசாயன மருந்து. (திருநெல்.), (கூட), (நா)

களங்கூட்டு மடடபசூமர""ர *வி.* விளக்கமாற்றால் கதிரடித்த நெற்களத்தைக் கூட்டுதல்.

களஞ்சியம் மப"ய;உலைடஅ *பெ.* வீட்டின் உள்பக்க சுவரை ஒட்டி பலகை/செங்கற்கலால் தடுத்து தானியங்கள் சேமிப்பதற்காக அமைக்கப்படும் கொள்கலன். (தூ.) (படம்). **களிஞ்சம்** மபடஉ;உயஅ (தரு)

களத்துநெல்லு மபடபவவரநேடடர *பெ* நெற்கதிரிலிருந்து நெல்லைப் பிரித்தெடுக்கும் போது களத்தில் சிதறிகிடக்கும் நெல். (நா.)

களத்துமேடு மயடபவவரஅநந"ர *பெ* ஏரிநீர் பாய்ந்து சாகுபடி செய்யமுடியாத திடல்பகுதி. (வே)

களப்பாச்சிநெல் மபடப"பயஉஎநேட *பெ* (பார்க்க–கலப்பு கதிர்). (தரு)

களம் மபடடஅ *பெ.* 1) நெற்கதிரிலிருந்து நெல்மணியை பிரித்தெடுத்து சுத்தம் செய்யக் கூடிய இடம். (தே.), (நாக.), (நா.), (விரு.), (தூ.). 2) செடியிலிருந்து ஆட்ந்த கடலையைக் கொட்டுவதற்காக தரையில் சிரியதாக வட்ட வடிவில் சுத்தம் செய்த இடம். (பெ)

களா மபடபய *பெ* வெண்மையானப் பூக்களையும் சிவப்பு நிறக் காய் மற்றும் கருமை நிற பழங்களையும் உடைய முட்கள் நிறைந்த ஒரு மூலிகைச் செடி. (மூ.)

களி மபடட *பெ.* (வேலையாட்களுக்கு மதிய வேலையில் கொடுக்க) கம்பு மாவில் உருண்டை வடிவில் செய்யப்பட்ட ஒரு உணவு. (பார்க்க–கூழ்) (வே)

களிமண் மபடஅயர **களிமண்ணு** மபடஅயரே *பெ.* அதிக ஈரத்தன்மையை பிடித்து வைத்துள்ள ஒருவகை மண். (தரு), (வே)

களியாப்பாட்டம் மபடலைபப"பப""யஅ *பெ.* திருநெல்வேலி மாவட்டம் புளியாரப் பகுதி.யில் உள்ள தேவஸ்தானத்திற்கு சொந்தமான நிலத்தில் சாகுபடி செய்பவர்கள் விளைந்தாலும் விளையாவிட்டாலும் தேவஸ்தானத்திற்கு நெல் அளக்கவேண்டும் என்று விதிக்கப்பட்ட ஒரு ஒப்பந்தம். (திருநெல்)

களிவா[1] மயடளைபட *பெ.* அதிகநீர் பிடிப்புத் தன்மைக் கொண்ட சேற்றுப்பகுதி.

களிவா[2] மபடளைபப *பெ.* விரைவாக ஈரத்தன்மையை இழந்து வரும் ஒருவகை மண். (தரு).

களை மபடடஉ *வி.* தானிய விதைகளை நீரில் விட்டு அழுக்கு அகலுமாறு கைகளால் பிசைதல்.

மண்டுதல் அபஅ"ரவபட *தொ.பெ.* களை/பயிர் பழுதில்லாமல் (செடிகள்) முளைத்து வளருதல்.

மண்ணஉருவு அயவேரசரஎர *வி.* மரவள்ளிக் குச்சி சாயாமல் இருக்கவும், நன்றாகக் கிழங்கு விடவும், மண்வெட்டியால் (பக்கத்தில் இருக்கும்) மண இழுத்து செடிமீது அணைத்தல். (நா.)

மண்ண எறக்குதல் அபவேட நபகபமமரவபட *தொ.பெ.* பட்டத்தில் பதித்த கரும்பின் இருபக்கத்தில் உள்ள மண்ணை கொத்தி சரித்து விடுதல். (கூட.)

மண்ணணை அபவேடேடன *வி.* 1) மரவள்ளிக் கிழங்கின் செடிகள் நன்றாக கிழங்குவிடவும் வளரவும் அச்செடியின்

வேரருகே கொத்திவிட்டு மண்ணை அணைத்தல். (நா.). 2) கடலை செடியில் நிலக்கடலை தோன்றும் பருவத்தில் அச்செடியைச் சுற்றி மண்ணை அணைத்தல். (கட.)

மண்ணணைத்தல் அயவேபேவைவபட *தொ.பெ.* (பார்க்க–மண்ணு கட்டுதல்). **(நீ.)**

மண்ண புளிக்கவை அயவே 'ரடமைமபளயன *வி.* உழுது போட்ட சேற்றை நீர்கட்டி பதம் மாறச் செய்தல். (கட.)

மண்ணு அடி அபரே ய"ை *வி.* (நிலத்திற்கு உரமாக) குளத்தின் அடிவண்டலை நிலத்தில் கொட்டுதல். (தே.)

மண்ணுகட்டு அபரேமப""ர *வி.* (பார்க்க–மண்ணு அணைத்தல்). (தரு.)

மண்ணுகட்டுதல் அபரேமப""ாவயட *தொ.பெ.* கோஸ் செடிகள் போன்றவை சாயாமல் இருப்பதற்காக கொத்திவிட்டு செடியின் வேரைச் சுற்றி மண்ணணைத்தல். **(நீ.)**

மண்ணுதிண்ணுதல் அபரேவனரேவட *தொ.பெ.* மண்ணில் அறிமானம் ஏற்படுதல். (நா)

மண்ணுவை அபரேஎபட *வி.* வரப்பின் ஓரத்திலிருந்து நீர் கசியாதவாறு மண்ணினை அணைத்தல். (தரு.)

மண்தகரம் அயவேயமபசயஅ *பெ.* மேட்டுப் பகுதி நிலத்தில் உள்ள சேற்றை பள்ளமானப் பகுதியில் கொண்டு நிறப்பும் பொருட்டு பயன்படும் பெரிய தகரத் தட்டு. (தஞ்)

மண் தொம்ப அபவேவடிஅ'ப *பெ.* (பார்க்க–குதுரு). (கட.)

மண்புழு ஓரம் அபரேரட டசடஅ *பெ.* மண்புழுவினால் மக்கச் செய்த நாட்டு எரு. (தஞ்)

மண்வெட்டி அபஏநே""ை *பெ.* (மண்ணை வெட்டுவதற்குப் பயன்படும் வகையில்) காம்பு மரத்தாலும் வெட்டும் பகுதி இரும்புத் தகட்டாலும் செப்த ஒரு விவசாயக் கருவி. (ப.து.), (சிவ.). **மம்டி** அபஅ"ை (திருவ.), (வே.). (பாகம்–மமுட்டி மிண்ட, காவு) **(நீ.)**. (பாகம்–பொடங்கு, கொளச்சி, கை, மொவா). (திருநெல்.). **மம்பட்டி** அபஅ'ப""ை (தே.), (திருநெல்.). **(நீ.)**, (ப.து.), (தஞ்). (பார்க்க–கொளச்சி மம்முட்டி). (பாகம்: எல, காம்பு, பிடங்கு, கெளிச்சி/கொளஞ்சி) (தூ.). (பாகம்–கண, பூனு, ஆக்க, பொடங்கு) (சிவ.), (விரு.). **மம்புட்டி** அயஅ'ர""ை (தஞ்.), (திருச்.), (திருநெல்.). **மம்முட்டி** அபஅஅர""ை (பெ.). **சனுக்க** உபrமமப , **சனிக்கி** உடமைமை (தரு.). **நம்பட்டி** பேஅ'ய""ை (பாகம்: நுனி, பூனு, கண, அடி, எவ) (ராம.). **மமட்டி** அயஅய""ை (முத்து.). (பாகம்–1: பிச்சு, மமுட்டி, பூணு, பொறடி, பொறடிகணம், புடி). (பாகம்–2 (கல்.): பொடங்கு, பூணு, தகடு, புடி, காம்பு (ப.து.). **மமட்டி** அயஅய""ை (நா.). **கைகொட்டு** மயமைடி""ர (கட.). **மமுட்டி** அபஅர""ை (தே.), (வே.), (நாக.), (நா.), (கட.), (தஞ). **கொத்து** மிடிவவர (கட.) **வம்பட்டி** எயஅ'ட""ை (தூ.), (ப.து.)

மணகண்டநூத்து அபயேமயமே"பபேபவர *பெ.* மணப்பாங்கான நிலத்தில் முளைத்த நாற்று. (தஞ்.)

மணச்சாரி அபயேஉபயசன *பெ.* மணற்பாங்கான நிலப்பகுதி. (தஞ்.), (திருவா.), (திருச்.). **மணக்கால்** அபயேமமபபட (ப.து.). **மணல்காடு** அயபேடமபய"ர (நா). **மணல்சாரி** அபயேஉபயசன (விரு.)

மணப்பாரமாடு அபபே"டபசபஅபப"ர *பெ.* விவசாயத்திற்கு நல்ல உடலுழைப்பைத் தரக்கூடிய மணப்பாரை என்ற ஊரில் இருக்கக் கூடிய மாடு. (கட.). **மணப்பார** அபபே"டபசப (தஞ்)

மணலிக்கீரை அபயேடமைமனைசபன *பெ.* சதைப் பற்றான கரண்டி போன்ற இலைகளையும் தரையில் படர்ந்து வளரும் சிறு செடி. (மூ.)

மணி அபவே *பெ.* முற்றாத மிளகு செடியின் காய். **(நீ.)**

மணித்தக்காளி அபவேவடமமயபடன *பெ.* நீள் வட்ட வடிவ இலைகளையும் கோணல் மாணலான கிளைகளையும் கொண்டு வெண்ணிற பூங்கொத்து மற்றும் கருநீலப் பழங்களையும் உடைய முட்களற்ற குறுஞ்செடி. (மூ.)

மரக்கலப்ப அடசமமபய"ப *பெ.* உழும் கொழு இரும்பாலும் மற்றப்பகுதிகள் அனைத்தும் மரத்தாலும் செய்து மாடுகட்டி இழுக்கக்கூடிய ஒருவகை உழுகருவி. நாட்டுக்கலப்ப. 1) முழுவதும் மரத்தால் செய்து மாடுபூட்டி உழக்கூடிய ஒருவகை நாட்டுக்கலப்பை. (விரு.). 2) நிலத்தை உழுவதற்குப் பயன்படும் ஒருவகை நாட்டுக் கலப்பை. (பாகம்: மேலி(ழி), கள்ளாணிக் குச்சி, கலப்பக்கட்ட, கொண்டி, மேக்கால், தடி). (தே.). 3) உழும் கொழு இரும்பாலும் மற்றப்பகுதிகள் மரத்தாலும் செய்து மாடுபூட்டி உழக்கூடிய ஒருவகைக் கலப்பை. (ம.). 4) கம்பி போன்று கொழுவைப் பொருத்தி முழுவதும் மரத்தாடல் செய்து மாடுபூட்டி உழக்கூடிய ஒருவகை நாட்டுக் கலப்பை. (பாகம்: மேழி, கைப்பிடி, ஏர்கலப்ப, கொண்டி, ஏர்கா, தொடகயிறு, மொகத்தடி, மல்லுமுடிச்சி, தும்பு). (ப.து.). 5) மாட்டைப் பூட்டி நிலத்தை உழுவதற்குப் பயன்படும் மரத்தாலானக் கருவி. (நா.). 6) மரத்தால் செய்யப்பட்டு மாடு பூட்டி மண்ணை உழக்கூடிய கலப்பை. (பாகம்–மோழி, ஊத்தாணி, ஏர்கால், ஆட்பு, குத்தி, இரும்புகொலு, நோக்கால், ஏர்காமுடிச்சி.). (திருநெல்.). 7) மரத்தால் செட்து மாடுகட்டி, உழுவதற்குப் பயன்படும் ஒருவகை நாட்டுக் கலப்பை. (பாகம்: பாவு, மேழி, கலப்ப, ஏர்கா, எடக்கா, மேக்கா, கொழு, தும்பு, குத்தி) (தூ.). 8) மாடு பூட்டி இழுத்து நிலத்தை உழுவதற்குப் பயன்படும் மரத்தால் செய்யப்பட்ட ஒருவகை உழு கருவி. (பாகம்–மோழி, ஆப்பு, தடி, வாலு–திருப்பு., கொண்டி, ஆணி). (தஞ்.), (கட.), (பெ.), (தரு.), (நாக.). **நாட்டுகலப்ப** பேப""ாமயடப"ய (பாகம்–மேழி, அள்ளாணி, மேலாப்பு, கீழாப்பு, ஏர்கா, கொழுவு, கயிறுவடம், நோக்கா, நோக்கா

தொள, பூட்டாங்கயிறு, ஒட்டுமேழி.) (திருநெல்.), (கட.). (பார்க்க–மரகலப்ப). (வே. புழுதிகலப்ப 'ரடரவமையடய"ப (பாகம்–ஏர்கால், மேலி, ஆப்பு, கொண்டி, கொலு, கலப்ப, நெகத்தடி, அடிவடம், எதுவடம்) / மரக்கலப்ப அபசபமமடடட"ப (சிவ.)

மரக்கா அபசபமமட பெ. 1) தானியங்கள் அளக்கக்கூடிய நான்குபடி கொள்ளவு கொண்ட ஒருவகை முகத்தலளவைக் கருவி. 2) அக்கருவி கொள்ளவு கொண்ட ஓர் அளவு (ஜமீன்). (தூ.), (தே.), (சிவ.), (விரு.), (புது.), (கட.), (தஞ்.), (திருவ.), (தஞ்.), (திருநெல்.) (புலி.), (ம.). மரைக்கா அடசயமைமபய (நா.). 2) 5 சேர்/7தி கிலோ கொள்ளவு கொண்ட ஓர் கருவி. (2) அக்கருவி கொள்ளளவு கொண்ட ஓர் அளவை. (வே.). 3) 4தி சேர் கொள்ளவு கொண்ட ஓர் அளவு கருவி. (2) அக்கருவி கொள்ளவு கொண்ட ஓர் அளவை. (வே.), 4) நான்கு பக்கா கொள்ளவு கொண்ட ஒரு முகத்தலளவைக் கருவி (ஸ்ரீவை). (தூ.). 5) ஆறுபடி தானியங்கள் கொள்ளவு கொண்ட ஒரு முகத்தளளவைக் கருவி. 2) அக்கருவி கொள்ளவு கொண்ட ஓர் அளவை. (ராம.)

மரக்காரை அடசபமமமடசயட பெ. கரும்பச்சை நிறமான இலைகளையும் இலைக் கோணங்களில் முட்களும் கொண்டு வெள்ளை நிறமலர் மற்றும் மஞ்சள் நிற கனிகளையும் தரக்கூடிய குறுஞ்செடி. (மூ.)

மரகுருது அடசடமரசரவர பெ. அடுக்கடுக்காக செய்து கோர்க்கப்பட்ட நாற்சதுரமுள்ள பெரிய மரப்பெட்டி. (தஞ்.)

மரசால் அபசயஉபயட பெ. தானியங்கள் சேமிக்கக் கூடிய ஒருவகை கொள்கலன். (சுவரின் ஓரம் நாலுடக்கமும் அடைத்து உள்ளே தானியங்களை கொட்டி மேல் டக்கம் மண்பூசி மெழுகி வைத்து பாதுகாக்கும் ஒருமுறை). (ராம.)

மரசால் அபசயஉபயட பெ. தானியங்கள் சேமிப்பதற்கு வீட்டில் செங்கற்களால் தடுத்துவைக்கப்பட்டத் தனி அறை. (விரு.)

மரத்தட்டு அடசயவவவப""ர பெ. உணவு உண்பதற்குப் பயன்படும் மரத்தால் செய்யப்பட்டத் தட்டு. (ம.)

மரத்தொம்ப அபசயவவடஅ'ய பெ. மூங்கில் குச்சியால் வட்டவடிவில் பிண்ணப்பட்ட குதூரு போன்ற தானியங்கள் சேமிப்புக் கலன். (கட.)

மரப்படி அபசட"ப"ை பெ. தானியங்கள் அளக்க மரத்தால் செப்யப்பட்டப் படி. (புது.)

மரப்பெரம்பு அபசப"நசயஅ'ர பெ. சேற்றினை சமப்படுத்துவதற்கு மரத்தால் செய்யப்பட்ட ஒரு தடித்தப் பலகை. (தரு.)

மரம் அடிச்சிபோடு அபசபஅ ய"உைஉ கை'டிடி"ர வி. பரம்பு பலகையால் சேற்றினை சமப்படுத்தி வைத்தல். (தூ.).

மரம்வெட்டுகத்தி அபசடஅஎந""ரமடவவை பெ. மரத்தை வெட்டக்கூடிய வளைவில்லாத ஒருவகை கத்தி. (நீ.)

மரமேரி அபசடஅநநசை பெ. தென்னை, பனை, பாக்கு போன்ற மரங்களில் ஏறக்கூடிய ஆள். (நா.)

REFERENCES

[1] H. Bejoint, *Modern Lexicography: An Introduction*, Oxford Oxford University Press, pp 6.32 (Ch.1), 2000

[2] D A. Cruse, *Lexical Semantics*, Oxford: Oxford University Press, 1986

[3] R.R K. Hartmann, *Lexicography: Principles and Practice*, London: Academic Press, 1983

[4] R.R.K. Hartmann, and G. James, *Dictionary of lexicography*, London Routledge, 1998

[5] F. Hausmann and HE Wiegand, 'Component parts and structures of General Monolingual Dictionaries: A Survey', in Franz Josef Hausmann and Oskar Reichmann (eds), *An International Encyclopedia of Lexicography*, Berlin· Walter de Gruyter, pp 328-359, 1989

[6] R.F. Ilson, 'Lexicography', in Asher (ed), *The Encyclopedia of Linguistics*, London Routledge, pp 29 I-298, 1996

[7] G. James, *Colporul· History of Tamil dictionaries*, Madras: Cre-A Publications, 2000

[8] B A. Kipfer, 'Work book on lexicography', in R R.K. Hartmann (ed), *Exeter Linguistic Studies*, Exeter· University of Exeter, Vol.8, 1984

[9] S.I Landau, *Dictionaries*. The art and craft of lexicography, New York Scribner, 1989

[10] Y. Malkiel, 'A typological classification of dictionaries on the basis of the distinctive features in Householder and Saporta (eds), *Problems in Lexicography*, pp. 3-24, 1967

[11] G. Matore, *'Hisloire DES dictionaries francais (Paris: Larousse)*, pp,1 8-22 (Ch. I), 1968.

[12] L.E. New Well, *Hand book on lexicography* Manila Linguistic Society of the Philippines, 1995

[13] T.A. Sebeok, 'Materials for a typology of dictionaries' *Lingua*, 1:363-374, 1962

[14] R.A. Singh, *An introduction to lexicography* Mysore. CIIL, 1982

[15] Bo Svensen, *Practical lexicography*, Oxford: Oxford University Press, 1993

[16] L. Zgusta, *Manual of lexicography*, Mouton: The Hague, 1971

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt of India.

8.16 Rule-based Machine Translation System using Indian Logic for Discourse Texts

Kommaluri Vijayanand

Abstract—**Communication is said to be taken place when the message delivered is received effectively at the receivers end. The communication channel and mode of communication are least signicant than delivering the message to the targeted audience. Several texts appear in day to day life which carries messages to its targeted audience in the form of advertisements, billboards, posters, banners etc. Though the message delivered is precise in its form, such messages are received by the public effectively. No linguistic information is found in such texts. The mechanism applied by the people to understand such texts is the 'Reasoning' and 'Interpretation'. Representing such reasoning and interpretation capabilities to the machine is one of the challenges to Articial Intelligence Researcher. Among Indian philosophers who worked on Logic, Acharya Dignaga is a signicant logician, also called a 'Fighting Bull'. The present paper is aimed to present the issues in Critical Discourse Analysis (CDA) and Knowledge representation using Dignagas Nyaya-pravesa, towards Machine Learning along with the implementation details.**

Index Terms—**Word Sense Disambiguation, Indian Logic, Discourse Texts, Machine Learning, Reasoning, Knowledge Representation, Critical Discourse Analysis, Perception, Inference.**

I. Introduction

ARTIFICIAL Intelligence (AI) is a dynamic area of research, that provide scope for transforming an idiot (the Computer) into intelligent. In doing so, a computer scientist need to equip the computer with knowledge that is laid down in the society in various forms. Till now, several Machine Translation (MT) systems were reported in India and abroad using various approaches and techniques These systems are fed with knowledge in the form of electronic dictionaries, the example-base with bi-lingual, tri-lingual and multi-lingual parallel corpus, the rule-base, etc. Question-Answering Systems, Interactive dialogue systems, Online Hotel and Railway reservation systems, etc. were also reported which rely upon AI that is fed in different forms. Most of the input, received in the form of text or speech from the general public to the system, does not meet the linguistic rules and requirements. The MT systems are prone to generate erroneous output due to ambiguity, when fed with such inputs. Under such circumstances, an urgent need to equip the computer to think on its own is identifed. This is possible when the computer can judge based on logic. Towards such initiation, an attempt is made in this paper to identify the possibility of knowledge representation using Indian Logic towards handing the discourse texts.

Logic is delivered by several Indian philosophers and Logicians, dated back from 1500 B.C. in the form of Vedas Among such great philosophers, Ācārya Dignāga is a signicant and earliest Buddhist writer on Logic, also called a "Fighting Bull" or a "Bull in discussion". He was regarded as a father of Mediaeval Logic. The branch of learning was paid attention differentiated from general philosophy began in 450 A.D Dignaga, in his writing, divided valid knowledge into six components, viz. Perception, Inference for ones' own self, Inference for the sake of others, Reason and example, Negation of the opposite, and Analogue. Dignāga stated two pramānās (means of valid knowledge), namely, the perception and inference. These components of knowledge are further elaborated in section I-A.

The development of precise frameworks of discourse interpretation has been hampered due to the lack of deeper understanding of the dependencies between different discourse units. This problem was initially visualised by Acharya R. M. K. Sinha and solved to a great extent with ANGLAB HARTI approach in Indian context [5] for Hinglish that accepts input in the form of a variant discourse texts. The problem of discourse language attracted more attention while working with the news texts, where the headlines contain purely discourse sentences [8]. A number of strong constraints have been proposed that restrict the sequencing and attaching of segments at various descriptive levels, as well as the interpretation of their interrelations. The discourse texts are to be analyzed in terms of psychological, social and cultural practices of the group unto the society level. The utterance level of the text need to be analyzed that specifie the importance and signifance of the event. To analyze these signifcant aspects of the discourse text a special kind of annoated corpora is essential apart from logic. However, discourse texts can be analyzed at a wide range of domains rather than in general form of language utterances.

During the literature survey to work on this present area, it is noted that the theoritical research work towards Critical Discourse Analysis (CDA) is in its infancy. So far, to the best of my knowledge there is no Machine Translation system working towards CDA for the language English itself. Thus an attempt was made to explore the possiblity of implementing CDA for the language English at this stage. The reason behind such idea is the availability of adequate lexical resources and knowledge in English

This paper is intended to present a theoritical framework, aimed towards investigating the possibility of applying Indian Logic preached by Dignaga, for handling discourse sentences during MT Section I A presents the theory of Dignāga,

Kommaluri Vijayanand is with the Department of Computer Science Ramanujan School of Mathematics and Computer Science Pondicherry University, Puducherry-605014, India (e-mail.kvixs@yahoo.co.in)

emphasized on various forms of valid knowledge, followed by the study on CDA that supports solution to deal with the discourse sentences that appear in texts during MT n section II. Section III presents the architecture for applying Logic to the machine. Section IV exemplifie various lexical resources used and their implementation details, followed by the conclusions in section V.

A. Knowledge and Analysis

The Nyāya praveśa is an excellent work on Logic by Dignāga. The core theme of this volume [7] is stated briefl as:

> Demonstration and refutation together with fallacies are useful in arguing with others; and Perception and Inference together with their fallacies are useful for self understanding; seeing these I compile this Sastra

B. Sources of Knowledge

Perception and Inference are the two kinds of knowledge for one's own self. Perception is the knowledge derived through the senses whereas Inference is the knowledge of objects derived through a mark. Perception of a thing consists of the knowledge of its individual characteristics alone. Knowledge derived through inference is general and can be well expressed by name, genus, etc., whereas, that derived through Perception is particular and is capable of being properly communicated to others by name, genus, etc.

Inference for one's own self is defined as the knowledge of a thing derived through its mark or sign (middle term) of three characters, viz. Effect, Identity, and Non-perception. An inference for the sake of others takes place when a person demonstrates to others the conclusion drawn by him through inference for one's self. The affirmative reason signifies that the thing signifie by its invariably accompanied by the thing signifed by the predicate. Another source of knowledge is Apoha, an entity being the negation of its opposite.

C. Reasoning

The literature against Logic is found in Vedas (composed between 1500 B.C. and 600 B.C.) which are regarded to be the oldest records not only of India but of the whole Aryan world. The term "Nyāya" in the sense of Logic does not appear to have been used in literature before the first century A D Panini (about 350 B.C.) did not know the word "Nyaya" n the sense of Logic [1] , and even Patanjali (about 150 B.C.) does not seem to have been conversant with the word.

1) Nyaya śastra: Nyaya śastra is the science of true reasoning. According to [6], Nyaya is defned as an examination of objects by evidences. He takes evidences to signify a syllogism which consists of a 'proposition' based on verbal testimony, a 'reason' based on inference, an 'example' based on perception, an 'application' based on comparision, and a 'conclusion' based on all the previous four. Vatsyayana (about 400 A.D.) uses the expression "parama ny āya"for the conclusion (niga mana) which combines in itself all the fve parts of a syllogism.

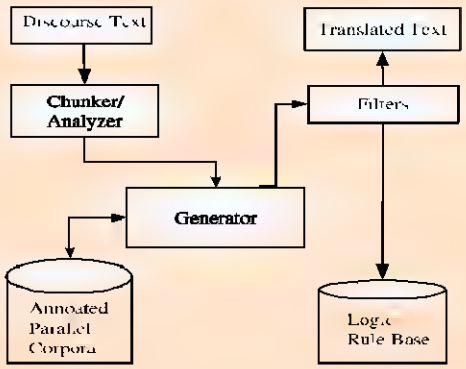


Fig. 1. Architectural Design for Discourse Translator

Dignaga (about 500 A.D.) explicitly mentions the fve parts or members of a syllogism as Nyāyāvayava. In the Sabhaparva of the Mahabharata we fnd Narada was an expert in Nyaya śastra. Narada is said to be clever as he could ascertain the validity and invalidity of a speech of fve parts.

2) Nyaya praveśa: Reasoning, according to the Nyaya praveśa, is carried out by means of a major term, a middle term, and two examples The minor term also called the subject, the major term also called the predicate, and the middle term called the reason, combined together with two examples form the reasoning. A minor term and a major term combined together form a preposition, e.g., The hill (minor term) is ery (major term). The characteristics of the middle term are:

a) The whole of the minor term must be connected with the middle term.
b) All things denoted by the middle term must be homogeneous with the things denoted by the major term.
c) None of the things are heterogeneous from the major term must be a thing denoted by the middle term.

Let us assume subject to be , reason to be R, and predicate to be P. Now, we can generalize the characteristics mentioned above as shown in example 1. The negative aspect of the middle term only confrms the truth conveyed by one of the positive aspects, viz. All R is P. Hence, we can omit the negative aspects and exhibit the positive aspects as depicted in example 2. However, R and P may be taken in whole extent or partially. Hence, the positive aspects mentioned above may be fully exhibited as given in example 3.

(1) **All** is R.
All R is P.
No R is **non-**P.

(2) **All** is R.
All R is P.

(3) **All** is **all** R.
All is **some** P.

(4) **All** R is **all** P.
All R is **some** P.

Combining the aspects 3 and 4 together, we find that a syllogism may be of any one of the forms shown in

examples 5, 6, 7 and 8. Therefore, we can conclude that All is all P, and All is some P. Another interesting aspect is the relative extension of the middle term and major term. They show that the middle term is universally, invariably, or inseperably connected with the major term. This is called Vyapti in Sanskrit.

(5) **All** is **all** R (Conclusion):
Because **All** is **all** R,
All R is **all** P.

(6) **All** is **some** P (Conclusion):
Because **All** is **all** R,
All R is **some** P.

(7) **All** is **some** P (Conclusion):
Because **All** is **some** R,
All R is **all** P.

(8) **All** is **some** P (Conclusion):
Because **All** is **some** R,
All R is **some** P.

(9) The hill is fier.
Because it has smoke,
All that has smoke is fery as a kitchen.

(10) Whatever is not fier has no smoke as a lake.

An example can be converted into a universal proposition i.e., Vyapti which stand to each other in the causal relation. The example 9, is homogeneous. The heterogeneous example may be laid down as shown in example 10.

II Critical Discourse Analysis

Today, language and meaning are in some way social constructs. Emphasis on both the structure and the social context of media texts can provide a solution. CDA is leading to the development of a different approach to understanding media messages. The basics of a text consist of syntax and lexicon; its grammar, morphology, phonology, and semantics. However, the understanding of grammar and lexicon does not constitute the understanding of text. The comprehension of meaning lies not in the text itself, but in the complex interaction between the author's/receptor's intent and his/her performative ability to encode/decode that intent. CDA provides an opportunity to adopt a social perspective in the cross-cultural study of media texts. Texts are selected and organized syntactic forms whose "content-structure" reflect the ideological organization of a particular area of social life. The language of media, as a particular style of discourse, is a complex blend of national, social, economic, and linguistic traditions which work in tandem with audience expectations. Discourse analysis employs the term in two broad categories of use [4]:

1) Discourse as an abstract noun denoting language in use as a social practice with particular emphasis on larger units such as paragraphs, utterances, whole texts or genres.
2) Discourse as a countable noun denoting a 'practice not just of representing the world, but of signifying the world constituting and constructing the world in meaning'.

A. Interpretation

Language cannot be seperated from interpretation. Analysis need to be sensitive to their own interpretative tendencies and social reasons for them. Textual interpretation is phychological that is derived by the intelligence of a person. Such intelligence is applied to decode the text with the background information in understanding the text. Not only do different types of text require different ways of reading, but the same can also be read in different ways to generate different meanings. The relationship of the lexicon to the social context of the utterance can be thought of as exemplifying the way in which codifed sign systems in general (verbal, visual, behavioural) are rendered meaningful only in relationship to the social structures which constitute them [2].

The texts that appear as discourse deal with the social structure of the society and need insight into the social and cultural aspects of the society. Recently, I had come across a billboard in my University that attracted me. That is a billboard afied by the eco-lovers of the society reproduced in the fgure 2 and 3. The text appeared in the billboard contains text and images. As a Computational Linguist, I fnd only one valid sentence, i.e., "Water is Life". The next phrase

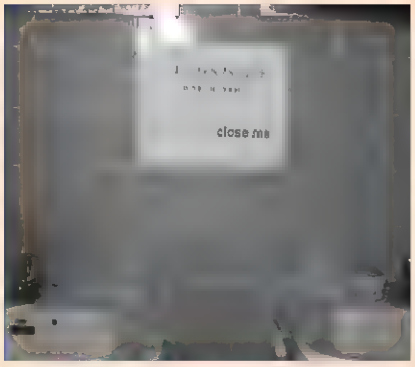


Fig. 2 Billboard at wash basin

Fig. 3. Billboard at to let

contains "Watch every Drop", where Drop is an image. The link between this sentence and the previous one is Water The reader receives the knowledge from the previous sentence to understand the image of a "Drop of water". Rest of the billboard requires some 'common sense' to understand. In addition to understand the message delivered, the location of the appearence of the billboard is very important and considerable. Though, the message appeared to be precise, it delivers a message of paragraph to the audience. When we compare the figure 2 and 3, the meaning one can derive can be same or different. If the reader do not possess any additional knowledge the meaning shall be the 'same'.

III The Architecture

Based on the problems posed and analysis emphasized on understanding towards translating the discourse texts, we had designed an architecture. The core components of the proposed architecture are the Discourse analyzer, the Generator, and the Filters. The input to the system are discourse texts which commonly appear in news headlines. The chunker or the analyzer receives these input texts and tokenize them according to the pre-defied constraints that appear in the text like newline characters, size of fonts, images, etc. Based on these tokens, the Generator shall retreive all possible translations, based on a specific approach viz., Rule based, Statistical based, Example based or Hybrid based approaches. The outputs from the Generator are received by the Filters, who shall apply the logic and context based knowledge to discard the irrelevant translations. Thus, the sentences after filtering shall be the actual translations for the discourse texts.

A The Analyzer

The chunker or sentence analyzer acts based upon the important aspects of the text received. The analysis of the text shall be composed with the Prosody, Cohesion, Discourse organization, Con textualization signals, and Thematic organization. The system shall differentiate between ordinary pauses like comma, semicolon, hyphen, colon, etc. and the pregnant pause The core theme is that, machine needs to identify the perceptiveness and authority of the writer. Verbal indicator of the prosodic features of stress is to be identified This can be accomplished by identifing the words like long, long been suspected, barely, far, absolutely etc. This is an important task in identifig the writer's argument towards his authority by a form of reiteration. Cohesive links are to be recognized which help in finding out the thread that tied the language and sense together. These cohesive links are conjunctions, pronouns, demonstratives, ellipses, adverbs and repeated words and phrases.

The rule base contains a set of rules that will govern the behavior of the filters during the filtering process. Those strategies that can be adopted during implementation are as follow:

1) Choose the rules which best approximates the knowledge.
2) Design the rule to model the knowledge.
3) Learn the rules from examples generated successfully.

Applying these principles we implemented the machine translation system as explained in the section IV. The rule base shall represent the domain specific knowledge

B. Knowledge Representation

The logic that was discussed in section I C can be represented using the Generalized Quantiers (GQ). A GQ can be define as: for a given set $\mathbf{t}$, a GQ on $\mathbf{t}$ is a relation among subsets of relations on $\mathbf{t}$. For example, When A universe $\mathbf{t}$ is fied, we use Q as a variable over GQs, A, B, ... as variables over sets, and write $Q(A, B)$ to indicate that A, B belong to the extension of Q. That is, they are in the relation denoted by Q. However, few quantifier are similar to usual logic constants and behave independently towards the context.

$\mathbf{all} - \{A, B\}$
$\mathbf{some} - \{\mathbf{at\ least\ n}\}$
$\mathbf{no} = \{\mathbf{at\ most\ n}\}$

IV. Implementation Details

Towards implementing the system for CDA, what is essential is the domain specific knowledge. For example, the text that appeared in figur 3 would be perceived by a reader not to waste the water. Such understanding will be computed mentally with the knowledge that is acquired and stored in one's memory. Therefore, unless the system/human possess the domain specific knowledge, it is impossible to understand the meaning. The domain specific knowledge that is involved in CDA shall be the knowledge about the culture, behavior, habits etc. The discourse texts do not contain the complete sentence, complete meaning, complete message. Now a days discourse texts are becoming more popular in advertising. An example can be the advertisements for liquor brands. The message is directed just to the alcohol consumers bypassing the minors. Hence, we need to train the system with a kind of knowledge to be aware of the general knowledge

A. Lexical Resources and their Representation

Such training is essential to transform it to be intelligent, we need lexical resources where the system shall rely upon. Hence, we had attempted to teach the system with the knowledge base created using a set of rules. These rules are collected from the reasoning exercises and trained the system using Prolog. The examples given in I1, I2, I3 and I4 are some of the forms of rules that are acting as knowledge base to the system, which can guide the system to conclude after its reasoning process. In all circumstances, the rule I1 shall treat that **All human are mortal** or **All men are human**. Like wise the experiment was successful to generate the conclusions for any given input with the help of these rules The knowledge base can be improved to any extent according to the users requirement. However, at this stage the experiment was carried out to the extent of predicate logic with $\forall$ construct.

(I1) **mortal**(X)**:- human**(X)
human(men).

(I2) **tasty**(X)**:- sweet**(X)
sweet(rasagolla).

(13) **power** X):- **rangers** X)
rangers(spd).

(14) **good** X):- **faithful** X)
faithful(dogs).

B. Methodology

The system was implemented using the Rule based method ology where the input is tokenized and transfered to match with the knowledge base. For every given successful match it returns the conclusions. The system will form a predicate for every set of rules that are stored in the knowledge-base. The reasoning is carried out by applying predicate logic on the rules. Thus the system shall derive conclusions for a given pair of predicate sentences. The following example 15 narrates the reasoning for the construct. . Interface to Prolog was provided using Visual C++ for creating work space. Some of the screen shots are given in the appendix.

(15) $\forall X, X \;\; s \;\; M \;\; f \;\; X \;\; s \; a \;\; .$

V. Conclusion

The paper attempted to present several issues in discourse analysis along with Indian Logic that may be used towards knowledge acquisition. Towards handling various texts in In dian languages. Implementation of this system shall answer to several unsolved prob.ems in the area of Machine Translation While implementing the present system, we had used the rules in general instead of domain specifc knowledge. Though we had listed several general application areas of CDA, still this problem shall be useful towards computational requirements like on line MT during chatting and conferencing As media is the powerful channel that deliver information to large masses, operationally one yeild maximum throughput from the MT systems when delivered with minimum number of limitations. It was a major diffcult task to analyze and design the present MT system due to its multidisciplinary nature. We were successful in implementing the present system with minimum utilization of resources and producing the initial results. The future enhancements to the system are to extend the work to deal with the predicate logic for **some of**, **some of** and **for all** During the investigation of the present system it is observed that implementing the predicate logic **not** is impossible. Eloborate study is necessary to apply the principles and implementation details discussed towards Indian Languages

Appendix A
Screen Shots of the System

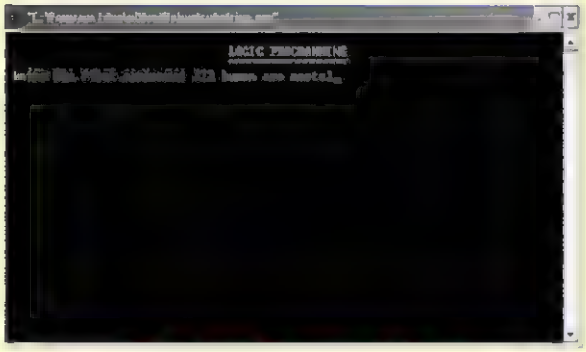


Fig 4. S reen Shot to enter the input

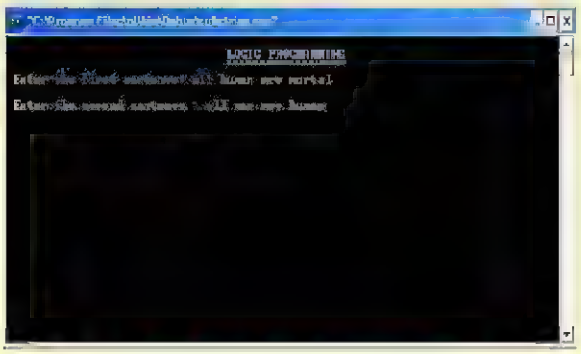


Fig. 5 Screen Shot after a pair of input

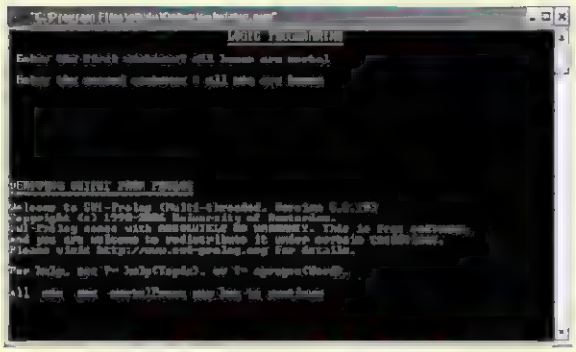


Fig. 6 Screen Shot to display the result

Acknowledgment

The author would like to dedicate this paper to his parents who were school teachers and fille their son with teaching interest. The author would like to thank Prof. S Kuppuswami, for his constant support extended and his colleagues of the Department.

References

[1] Goldstucker, "Panini," p. 157
[2] T. Loeke, Critical Discourse Analysis. London Continuum International Publishing Group, 2004.
[3] C. Moraga, E. Trillas, and S. Guadarrama, "Multiple valued logic and artificial intelligence fundamentals of fuzzy control revisited," Journal of Artificial Intelligence Review, vol. 20, pp. 169–197, 2003
[4] A. Pennycook, Critical Applied Linguistics: A critical introduction Mahwah, NJ: Lawrence Erlbaum Associates, 2001
[5] R. M. K. Sinha and A. Thakur, "Machine translation of bi-lingual hindi-english (hinglish) text," in Proceedings of the 10th Machine Translation Summit (MT Summit X), Phuket, Thailand, September 13-15, 2005
[6] Vatsyayana, Nyaya-bhasya, 1-1-1
[7] S. C. Vidyabhusana, A History of Indian Logic. New Delhi, India: Motilal Banarsidass, 1920
[8] K. Vijayanand and R. Subramanian, "Anuvadhi: An automatic example-based machine translation system for bengali into assamese and oriya" in Proceedings of the First National Symposium on Modeling and Shallow Parsing, Indian Institute of Technology, Bombay, April 2-4, 2006, pp. 267-281
[9] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," Journal of Artificial Intelligence Review, vol. 18, pp. 77-95, 2002
[10] D. Westerstahl, Handbook of Philosophical Logic. Dordrecht: Reidel Publishing Company, 1989

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.17 Art of Hindi Dictionary Making: An Historical Exploration

Ravikant Shrama, IndLinux

Abstract **Now that we are setting out to build new digital lexicons, it is perhaps useful to revisit the ways in which modern dictionaries were built in North India from the 19th century onwards. My paper will examine the shifting ambitions, cultural assumptions, politico-linguistic, lexical methodologies and usage of resources in the process of making various kinds of corpora that we have today. What has been the degree of community participation, if any, in projects that are by their very nature mammoth. How much have we really travelled from the days of Fallon and Platt through the age of Nagri Pracharini Sabha to that of Hardev Bahri and McGregor? Do we really see a shift in the post-independence period, especially after the setting up of the Rashtrabhasha Vibhags and expert bodies like CSTT? What relationship do these experts have with the languages of the society? What are their dreams and ambitions? Further, do we have anything to learn from the projects of the past? Or the art of digital dictionary making is an exercise entirely different from the print experience? Does it help to think in terms of computers on the net as a unique, interactive mass media? If so, what are its implications for dictionary making now?**

A language is not learnt from Dictionaries.
- S W Fallon

A hundred-odd years are not much in the life of a language. Yet if one looks at the history of Hindi spanning over the last century and a quarter or so, one really marvels at the distance it has traveled. There can not be any doubt about its continued expansion and enrichment in certain registers and spheres within India and beyond. Yet there is something about it that worries any lover of the language. It has been the Rajbhasha for more than sixty years now, yet it remains powerless. We have spent crores on its capacity building yet it remains intellectually resourceless. Now, the moment of Globalisation has thrown up contradictory signals thanks to the media market, Hindi has acquired an unprecedented visibility, its newspapers and magazines are clocking all time high circulation, you have all kinds of technical and financial programmes being presented at the ever busy Hindi TV and Radio FM channels and for the first time it appears Hindi has taken off from the literary stronghold it had encased itself in But if we do a reality check in terms of hardware resources, the state of dictionaries in the language is a good metaphor for the state of Hindi in general. This paper therefore tries to excavate the history of Kosh-making in modern Hindi with a clear agenda: Is there something in this history that can teach us? Are we in a position to learn before it is too late?

Ravikant Shrama is with Sarai-CSDS New Delhi (e-mail: ravikant@sarai.net)

To be sure Kosh-making is an age-old practice in India and its history has produced any number of good grammarians and lexicographers - Amarkosha and Nighantu Shastras belong to ancient India. But all of that was in the distant past, the past of languages that are alas dead and gone from our midst What is relevant for us, especially in the context of Localization, is the living languages the languages we still speak and write and hope to localize digital machines in For this I would like to pick up the story from Nineteenth Century when colonial masters were trying to understand who speaks what so that they themselves could communicate with their diverse subjects. Although the making of Colonial Lexicons is almost as old as Colonialism's efforts to expand in India in the 18th century itself, let us start with a more definitive work published in 1879. I will spend some time talking about SW Fallon's A New Hindustani-English Dictionary, because it captures a fuzzy moment in the history of modern lexicography, and because Fallon himself spends great deal of time explaining to us his intent and purposes an act not too many people would sadly repeat in future.

So what does Fallon's Dictionary look like? It is a 1216-page thick long tome that uses four scripts, Arabic-Persian and Devanagri and of course Roman for English. It showcases 'the spoken and rustic mother tongue of the Hindi speaking people of India; the exhibition of the pure unadulterated language of women; and the illustration given of the use of words by means of examples selected from the every day speech of the people, and from their poetry, songs, proverbs, and other folklore. In his 'Preliminary Dissertation', Fallon expresses his dissatisfaction with the way linguists have understood language as a refined and pure repository of select users with the result that they have ended up describing and indexing languages on the basis of their religio-linguistic roots. He rejects both Pandit Hindi and Mullah Urdu and goes for the middling, inclusivist Hindustani written in two scripts. The illustrations provide an uncanny amalgamation of dohas and she'rs, ghazals and kavitaai, kajri, holi and Marsiya. A special emphasis is laid on the *Zanana boli* or *Rekhti* and the basis for that choice is explained thus: "The seclusion of native females in India has been the asylum of the true vernacular, as pure and simple as it is unaffected by the pedantries of of word-makers. it is also the soil in which the mother-tongue has its most natural development....Yet this true vernacular is not all confined to the narrow home in which it has been kept alive The inherent vitality of living speech, and the all-pervading influence of women on language, strenuously strive to restore the deposed natives of the soil to their rightful inheritance....

If we conclude on the basis of this description that Fallon was involved in a romantic project of excavating the past for its own sake, we would be terribly mistaken. He was trying to make a dictionary that could be used by everybody including the colonial officials who were struggling hard to come to terms with Indian languages. To dispel any such notion he gives us a whole comparison of scientific terms translated into Aabic, Sanskrit and Hindustani. The results are breathtakingly useful, even today·

Retarded Velocity	घटती चाल	मोतनाक़िस मिक़दार-ए-हराकृत	न्यूणमान गति प्रमाप्य
Valve	खुल मुंदनी		
Forces in equilibrium	तुले हुए ज़ोर	मुयुल-ए-मोतदिला:	तुल्यमान शक्ति
Rotatory Motion	चक्कर चाल	हरकते बज़ाई	चक्र गति
Perpendicular Line	खड़ी लकीर	आमूद	लंबक
Proportion	बराबर निस्बत	तनासुब	परस्पर संबंध
Right Angle	खड़ा कोना	जाविया-ए-क़ायमा:	सम कोण
Obtuse Angle	फैला कोना	ज़ाविया-ए-मुन्फरेजक	अधिक कोण
Acute Angle	सुकड़ा कोना	ज़ाविया-ए-हद्दा:	न्यून कोण
Parellel Lines	बीच बराबर लकीर	खुतूते-मोतवाज़ी	समान अंतर रेखा
Diagonal	अध काट	क़ुत्र	भुज कर्ण
Sine	सामने की नाप	जैब	भुजज्य
Cosine	साथ कोने की नाप	जैबे मुस्तावी	कोटिज्य
Tangent	छूती नाप	ज़िल/मोमास	स्पर्श रेखा
Secant	काटती नाप	सहम	छेदन रेखा

Just look at the simplicity of the so-called rustic usage: Exactly the opposite of the contrived language of modern lexicography which is made exclusively from using the roots and derivative rules mostly if not exclusively from Sanskrit. More on that later. For now let us stick to some other lexicological efforts in the Colonial North India. We are all familiar with the story of how Hindi itself became a mission in the anti-colonial struggle and people wrote in it to liberate themselves. What I hope to show here is that the exercise of lexicography was pretty much a communitarian project till quite late. For this, the next text I choose to highlight is one concerned with Standardization of Hindi. Anant Choudhary's well researched work on the making of Hindi Language Script and Grammar in the early 20th century (*Nagari Lipi aur Hindi Vartani*, Bihar Hindi Granth Akademi, 1973) tells us that the task of standardization of Hindi orthography and grammar was actually taken up on a big scale. People sent out open calls in magazines, and those who knew language and could write, wrote in until there was a raging debate that went on for years. Fixers of language came up with models and each notation and grammar rule was hotly debated. This is a good contrast to our day and age when we see linguistic diversity and controversy as problems and shy away from taking opinions, especially from the people who use language and therefore know how it works

My third exhibit is the Sankshipt Hindi Shabdsagar, compiled from the Ramchandra Verma original Kosh and published by Nagri Pracharini Sabha in 1933 The kosh boasted of *lokprayukta* words taken from *deshi* and *videshi* sources and various *bolis*. Go to any page of the dictionary and you will any number of words from diverse, words that have become port of people's lexicon in the true blue *Bhasha bahta neer* style

The thing I want to emphasize is that it was a multilingual world, in spite of the parallel movement against the Urdu language and Script. There was an ambition to develop a script and register that could cater to all kinds of needs. That is how you had magazines like Devanaagar in the early decades of the last century, short-lived in the first instance and then revived again in the 1950s, that published material in various Indian languages written in the Nagri script. Granted that these efforts were limited and few and far in-between, but what we need to take from them is the idea that closing doors on languages would not make your own very prosperous.

It is not as if things changed dramatically after the famous midnight of Independence. There were several academies at work and the task of resource building for Hindi was much more institutionally de-centralized than it is today. Hindustani Academy for example had stalwarts like Dheerendra Verma who inspired people like Amba Prasad Suman to compile a Compendium of Agricultural terms in Brajbhasha. The two-volume illustrated compilation remains a delight inasmuch as it covers all the important tools, technology, practices and actions associated with agrarian life. You have terms on masonry, carpentry, metallurgy, irrigation, weaving, dying, oil-pressing, animal husbandry, pottery, ironsmithryas well as those associated with agrarian trade. The lexicon went beyond its set goal of overtaking Grierson both in terms of quality and quantity, as in a number of words. Let us see who and it is really a who is who of Hindi - is saying what in the blurb of a book published in 1960.

1. हिंदी का क्षेत्र विशाल है. उसकी विशालता का रहस्य उसकी उपभाषाएँ हैं. निस्संदेह हिंदी की उपभाषाओं में ही उसकी प्रतिभा छिपी हुई है. प्रस्तुत खोज प्रबंध इस सत्य को स्पष्ट करता है तथा विद्वानों एवं भाषा-प्रेमियों का ध्यान उस असीम खजाने की ओर आकर्षित करता है, जिसका उपयोग यदि शीघ्र नहीं किया गया तो हिंदी का प्रकृत स्वरूप; उसका निजी स्वरूप विलुप्त हो जाएगा.

- विद्या भास्कर, मंत्री तथा कोषाध्यक्ष, हिंदुस्तानी एकेडेमी, इलाहाबाद, 1960.

2. मेरी निश्चित सम्मति है कि अलीगढ़ क्षेत्र की बोली के आधार पर 'कृषक जीवन संबंधी ब्रजभाषा शब्दावली' शीर्षक बृहत शोध-प्रबंध हिन्दी बोलियों की समृद्धि का ऐसा पक्का प्रमाण उपस्थित करता है जिसे देखकर हिन्दी की अभिव्यक्ति क्षमता के प्रति मन में नयी आस्था उत्पन्न होती है...हिन्दी के कल्याण के लिए यह ग्रंथ छपना ही चाहिए.

---डा. वासुदेवशरण अग्रवाल

3. जनता की बोलियों में तद्भव शब्द बहुत बड़ी संख्या में पाये जाते हैं. साहित्यिक हिन्दी में इनकी संख्या कम होती जा रही है, क्योंकि ये गँवारू समझे जाते हैं. वास्तव में ये असली हिन्दी-शब्द हैं और इनके प्रति विशेष ममता होनी चाहिए. कृष्ण की अपेक्षा कान्हा या कन्हैया हिन्दी का अधिक सच्चा शब्द है.

---धीरेन्द्र वर्मा, भूमिका से:

4. विविध कला-कौशलों तथा व्यावसायिक शिक्षा के क्षेत्र में पारिभाषिक शब्दों की समस्या को हल करने के लिए हमें एक दूसरी दिशा में भी खोज कार्य को प्रवर्तित करना है. किसानों, मज़दूरों तथा अन्य श्रमजीवियों की बोलचाल की भाषा में समाजशास्त्र, शिल्प तथा उद्योग-धंधों के बहुतेरे बढिया-बढिया शब्द मिलेंगे जो राष्ट्रभाषा की समृद्धि के पूरक हो सकते हैं. ऐसे शब्दों का सर्वे और संग्रह कराना परमावश्यक है; अन्यथा केवल अंग्रेजी की तालिका तैयार करके उनका पर्याय प्रस्तुत करते जाने की परिपाटी पर ही निर्भर करने से हम अपनी लोकभाषाओं के हजारों अर्थपूर्ण उपयोगी जीवित पारिभाषिक शब्दों से वंचित रह जाएँगे.

--- विश्वनाथ प्रसाद, भारतीय हिन्दी परिषद के दशम अधिवेशन, 1952, (आगरा) में 'हिन्दी गवेषणा और पाठ्यक्रम का पुन:संगठन' शीर्षक से दिए गए भाषण से उद्धृत.

What we have here is a series of recommendations for using living resources whether in the tadbhav form or borrowed from the usages in the many sub-languages of Hindi And these are no mean characters we are quoting they were the best people in the business. We also have similar sentiments from scholars like Hajariprasad Dwivedi and Rahul Sankrityayn who collected and wrote on the folk literature and culture of the Indian subcontinent and beyond.

But things changed slowly and surely for worse in the post-Independence India, at the 'moment of arrival' for Indian Nationalism. Institutionally, it happened with the making of expert bodies at the top, it happened with the Rajbhasha Vibahags in various central and state government institutions. But there was also a deeper social, political and philosophical reason for the kind of Hindi we thought was good Hindi had won its cause in the Constituent Assembly and the fact of political Partition prepared the ground for a larger linguistic separation in the remaining part of North India. There was a definitive campaign launched by Hindi politicians like Ravishankar Shukla and others against the use of middling language Hindustani in the government-owned mass media such as All India Radio. The advocates for the exclusivist and pure type of Hindi were actually a paranoid lot they feared any affinity with Urdu or the so-called other dialects and espoused a direct lineage with Sanskrit In this imagination, all the 'foreign' as well as rustic elements in a language had to be weaned out. And so the new Hindi that was constructed in the government manufactories was sought to be exclusively Sanskrit-based

This created a massive rift in the public body of language and knowledge and created a curious paradox those who worked with living people's languages acquired the status of antiquarians and those who

recycled a 'dead' language in the name of the 'language of the nation-state became its leaders But that was the official approach and their success was as far as it could go! In spite of the omnipresent hoardings and road-signs, people did not really take to the officialese like 'bhoomigat paidal paar path' and 'durghatna aashankit Kshetra' for they were fond of another kind of language, a mixed language, which for example, was used in the Hindi films. Films worked with a different logic of market and profit and access was their mantra. They were not driven by an ideology. If films did not speak to people they would reject them they would be declared flops at the Box Office. Look at the picture now. Where is the Official Hindi now? Can you find much similarity between what is being spoken 24X7 on these new channels and the one we had come to identify as doordarshan Hindi let's listen to Hindi in news rather than Hindi news variety? Perhaps not. Let us look at the Hindi content on the Internet, at the various wiki-like efforts and see if there is departure from purist cananization of Hindi literature. There definitely is

So that is the point really of this paper. Computers connected to the Internet are mass media The mass media has to speak to, with and through the people. It is the people who would use these tools. Experts are best deployed at the back end. To localize tools is a welcome idea but to present the interface in a dead language is worse. It is better left the way it is in English or we try and be generous and innovative in our resource-building exercises. Time to hear S W Fallon and Hindi Cinema. Time to de-stigmatize Doordarshan and Akashvai Hindi. Time also to take the public domain route so that users of computers and languages have a say in the way their machines talk to them. In other words, time to ask people to make tools in their language

ACKNOWLEDGMENT

I must thank the organisers and participants of LRIL 2007, especially Dr Alka Irani and Ms. Shashi Palekar for their hospitality and warmth. I also owe huge debt to Prof. Shahid Amin for pointing me towards Fallon a decade or so back. My colleagues on the the Indlinux Project Gora Mohanty, Guntupalli Karunakar and Ravishankar Shrivastav have worked tirelessly to take it to where it stands now

SELECT READINGS

1 S. W. Fallon, *A New Hindustani-English Dictionary*, Uttar Pradesh Urdu Akademi, Lucknow, 1955 Edn

2. Amba Prasad 'Suman', *Krishak Jeevan Sambandhi Brajbhasha Shabdavali, vols.1 and 2,* Hindustani Ekedemi, Allahabad, 1960

3. Alok Rai, *Hindi Nationalism. Tracts for the Times, #13*, Orient Longman, Delhi, 2001

4. Ramchandra Verma, *Sankshipt Hindi Shabdsaagar*, Nagri Pracharini Sabha, Kashi, 1981, Edn

5. Ramchandra Verma, *Manka Hindi Kosh*, Hindi Sahitya Sammelan, Prayag, 1953

6. I have also seen and consulted the various volumes prepared by CSTT

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.18 Lexicographic Traditions in India and Sanskrit

Malhar Kulkarni, IIT, Mumbai

Abstract **Lexicographic Tradition of Sanskrit (Vedic age, Classical age [Amarakosa, other 11 lexica, analysis, the concept of synonymy and homonymy], Modern age [Wörterbuch, Monier Williams, Encyclopaedic Dictionary of Sanskrit on historical principles])... Grammatical lexicon... Application of Panini's grammar to create new lexica)**

Index Terms **Dictionary, Lexicographic, Lexica, Panini, Sanskrit, Synonymy, Homonymy**

I. INTRODUCTION

THIS paper is aimed at two things- 1) to take an overview of the lexicographic tradition of Sanskrit, highlighting the main features, and 2) to present a model of Sanskrit grammar towards developing new Lexica of newly coined terms for the purpose of creation and use of scientific and technological terminologies. The lexicographic tradition of Sanskrit is divided under three heads - a) Descriptive, b) Historical and c) Etymological. We here, propose to give a brief overview of the first two classes. Amongst them, we can demarcate three stages, chronologically-the Vedic age, the classical age and the modern age.

II. VEDIC AGE

Preservation of Vedic literature was the biggest motivation for the Linguistic activity in India some approximately 2500 years ago. This activity resulted in segmenting Vedic sentences into words, and words into root-suffix components. In this activity were rooted morphological, phonological as well as morpho-phonological theories that were developed and applied by Panini and his tradition in the grammar of Sanskrit. The oldest known lexica of Sanskrit are entitled Nighantu (700BC) on which Yaska is believed to have written a commentary called Nirukta There are believed to be several works known as Nighantus. These lexica are domain specific. They have only Vedic words as Lexemes. We find references in the Pali literature to Nighantus as well. The Nighantus have arranged lexical material from the point of view of Synonymy as well as Homonymy. Yaska in his commentary tried to explain these word-forms from the etymological point of view.

III. CLASSICAL AGE

Researches in the field of Philology and Linguistics in the past two hundred years or so have made available to us many lexica in the printed form. We propose to take stock of some of these in the present section and to comment on their methodologies.

A. *Amarakośa*

The first and the foremost popular name of a lexicon in Sanskrit is that of Amarakośa written by Amarasimha before 6h century AD. Much has been said about this lexicon and its arrangement. In fact, in one of the most valuable modern contributions to the subject at hand, namely, "History of Sanskrit Lexicography" by M. M. Patkar (being the revised version of the Ph.D. dissertation submitted to the University of Mumbai), the author devotes entries 9-27 out of the total 106 to Amarakośa and the commentaries thereupon and the mention of nearly 40 commentaries on this work in the Catalogous Catalogorum, speaks volumes of the importance and the popularity of this lexicon.

B. *Arrangement of Amarakośa*

The name of this lexicon is "**Namaliṅg-anuśasana**", which means a work that deals with the lexemes and their genders. It is also known as **Trikaṇḍaśeṣa** and as the name suggests, it is divided into three chapters Each chapter is further divided into sub sections called vargas. In the 1st Chapter, also known as Svargadikaṇḍa, there are nine subsections- Svarga (heaven), vyomadig (Ether and space), kala (time), dhi (intellect), śaila (mountains), natya (dramaturgy), patalabhogi (nether world), naraka (hell), and vari (water). In the second chapter, known as bhumyadikaṇḍa, there are ten subsections Bhumi (earth), pura (city), śaila (mountains), vanauṣadhi (Forest herbs), simhadi (animal), nṛ (human), brahma (Brahmin), kṣatriya (warrior), vaiśya (trader) and śudra. The last chapter, known as samanyakaṇḍa, has five subsections- Viśeṣyanighnavarga (adjectives), samkirṇa (miscellaneous), nanartha ,homonyms), avyaya (indeclinables), liṅgadisamgraha (miscellaneous).

Apart from a small subsection of the third chapter, this entire lexicon is devoted to Synonyms. In this small subsection the words are arranged after the final consonants. The formal arrangement is the unique aspect of the arrangement of this lexicon. The gender of a particular word is demonstrated by actually using the word in that gender. In the introduction, the author has clarified the point. He says-

(1) prāyaśo rūpabhedena sāhacaryācca kutracit / strīpuṃnapuṃsakaṃ jñeyaṃ tadviśe-ṣavidheḥ kvacit //

(Normally, the gender of a word is distinct from the difference of the form. Sometimes that is to be ascertained from the association of the other synonymous words, sometimes however, it is to be ascertained from the word which it qualifies.) Thus for example-

(2) kaumodakī gadā khaḍgo nandakaḥ kaustubho maṇiḥ /

The formal difference between "gada" and "khaḍga" tells the gender difference between two

Malhar Kulkarni is with the Department of Humanities and Social Sciences Indian Institute of Technology Mumbai (e-mail: malhar@iitb.ac.in).

words. In the same line the gender of the words "nandakaḥ", etc. is ascertainable from the association of it with the other words in the line There is a line at the end of dhivarga of the first chapter-

(3) guṇe śuklādayaḥ puṃsi guṇiliṅgās tu tadvati /

(The words "sukla", etc which when denote the *quality*, are in masculine. However, when they denote the *qualified*, they take the gender of the *qualified*

(4) triliṅgyāṃ triṣviti padaṃ mithune tu dvayoriti/ niṣiddhaliṅgaṃ śeṣārthaṃ tvantāthādi na pūrvabhāk//

(The word "triṣu" [in all the three genders] is mentioned after the word which appears in all the three genders, the word "dvayoh" [in two genders] is added after the words which appear in masculine as well as in feminine The negation of a particular gender indicates that that word appears in the rest of the genders. The words, "tu", "anta" and "atha" added before, suggest the demarcation of the previous synonym. Thus-

(5) vyomaynaṃ vimāno 'strī

This indicates that the words vimāna and vyomayāna are not to be used in feminine. The instruction here points out that they can be used in other two genders.

(6) tikto 'mlaśca rasāḥ puṃsi tadvatsu ṣaḍamī triṣu /

The words tikta, amla, etc. are used in masculine gender when used to indicate the fluids. However, when used to indicate the entity possessing them they can be used in all the three genders. This is indicated by the word *triṣu* used in the line.

C. *Features of Amarakośa*

Following are some of the features of this lexicon:

- Arranged in the verse form.
- No alphabetical arrangement.
- Classification of words according to Synonymy.
- Classification of words hyponymy. Names of Gods are stated first and then synonyms of each subset of god are stated It records ISA relations
- Syntactic features of words are pointed out. (Ex- adjectival relation mentioned)
- The important feature is that words are recorded not in the form of stem but in the form of inflected forms. That does not deny the awareness of the type of the stem on the part of the Lexicographer. Some words are mentioned in big compounds

(7) śukla-śubhra-śuci-śveta-viśada-śyeta-pāṇḍarāḥ

Only stems are mentioned here. Thus it is noted that the processes of grammatical alteration of sounds, augmentation etc. are known to Amarasimha. Apart from Amarakośa several lexica are constructed; many of which are often quoted by traditional commentaries Some of them are

Vaijayanti of Yadavaprakasa
Halayudha of Halayudhabhatta
Abhidhanacintamaṇi of Hemacandra

D. *Other Lexica*

These lexica were constructed more or less in the style of Amarakosa. There were some other lexica available in this period. Let us look at some of them closely. They are -

1. Namamalika of Bhoja- 11th Century
2. Siddhasabdarṇava of Sahajakirti- 17th Century
3. śāradiyākhyanamamala of Harsakirti- 17th Century
4. Paryayasabdaratna of DhananjayaBhatta
5. Kosakalpataru
6. Nanartharatnamala of Irugapa Dandadhinatha- 14th Century
7. Nanarthamañjari of Raghava
8. Dharaṇikosa of Dharanidasa- 12th Century
9. sivakosa of SivadattaMisra
10. Ekārthanāmamālā-dvyakṣaranāmamālā of Saubhari
11. Paramanandiyanamamala of Makrandadasa.

They can be classified broadly into three groups

(i) Lexica of Synonyms- 1-4
(ii) Lexica of Homonyms- 5-10
(iii) Lexicon of both Synonyms and Homonyms- 11

E. *Arrangement of These Lexica*

It is worthwhile to note the system of arrangement as well as some of the features of some of these lexica. The lexica of Synonyms are normally arranged topic wise. Lexica of Homonyms are constructed in a somewhat different manner. Thus for example, in the "8" mentioned above, words are arranged after the final consonants and the number of syllables. The homonyms are as a system arranged in the quarter as well as in half verse or sometimes in the entire verse. In this the conjunct "kṣ" as a letter of alphabet comes after h.

Thus if the words ending is 'k', we have following verses-

(8) śloko yaśasi padye syāt lokas tu bhuvane jane /

(The word 'sloka is used in the sense of fame as well as a verse; the word 'loka', however, is used in the sense of world and people.)In this, it is noted that one word is treated only in a quarter of the verse. Amongst such cases words having only two

vowels, namely, sloka, loka, stoka, etc. appear first. Then come the words- anaka, sayaka and jambuka, etc. with three vowels, and then the words with four vowels- utkalikā, etc. The same procedure is followed everywhere.

Later on we find a verse half portion of which deals with one word Thus-

(9) śulkam ghaṭṭe vivāhārthe jāmātur gṛhyate ca yat /

(The word 'sulka' is used in the sense of a ghat, as well as in the sense of what is collected from the son-in-law for the sake of marriage.) In this half line of the verse only one word, namely, sulka is treated. Sometimes one word is treated as the entire verse Thus

(10) śikhā jvālā śikhā cūḍā śikhā śākhā śikhā śiphā /śikhā śikhaṇḍinām cüòā śikhā syād agramātrakām //

This verse is devoted entirely to the word çikhä and its meanings.

In "6", homonyms are arranged in the order of the last syllable and not of the last consonant. This lexicon is seen to be influenced by Southern vernaculars. Thus in this lexicon, words with one akñara are listed first then with two and three and then four. Amongst them again words are arranged according to the alphabet. The words kuèòha (kueóa in kannada meaning dwarf), praya (práya in kannada meaning age), saëëa (saëëa in kannada meaning small) and koëa (koëe in kannada meaning room) are used in this lexicon in the same sense in which they are used in kannada. Normally these traditional lexica are constructed in anuñoubh metre, however, "5" is constructed in more than 15 metres like vasantatilakà, çäline, mälne, etc

"9" is a specialized homonymous lexicon restricted to the names of plants, trees, and herbs that go to form the Materia Medica in the Ayuvedic system of Medicine and it records about 2860 principal and about 4860 words denoting the meaning thereof There are almost 44 names of authors as well as works in the field of Ayurveda and about 44 works and names of authors of general lexica are found quoted in this lexicon. This lexicon provides the names of the original homelands of the medicines mentioned. Thus, words like çriparree, gambhäre, kaóphalà, hirà, etc., which indicate names of the plants, are stated to be originated in the region of Kashmir. Kalinga is mentioned as the homeland of läigala, kuõaja, gaura, etc.

"10" is a very peculiar type of lexicon, perhaps unique of its kind. It mentions words of only one and two syllables. Thus-

(11) kuù påthvé kuù kucaù kulam kuù kåtyä bhürapi småtä /

This line records words with one vowel with k consonant at the beginning with their meanings They have not been recorded by the modern Lexica.

In "11", following features can be highlighted- a) the homonyms are arranged into groups according to the initial letters and the number of syllables in each letter. b) In the section of Synonymous words a desi word is also recorded. c) Vernacular phrases are tagged before each set of synonyms

Compound words are not used as far as homonyms are concerned. Cases for anusvara are not dealt with 'k ṅ' appears after 'h'. We find-

(12) jambhéro jambhälo jambho jambéro deçyasamgrahe /

(13) preìakhitändolite cäpi deçye hinïcitam iñyate //

In a nutshell we can summarize the features of the traditional lexica as follows. Features of this phase are-

- The alphabets were ordered systematically according to phonetic principles; to this corresponded the macrostructure of the lexicon. However, alphabetical order, as far as the beginning of the word is concerned, was not followed
- Lexicography was developed systematically. The stem (or root) was used as the lemma.
- A theory of apophony was devised.
- A theory of compounding was developed
- Dialect differences were noted

F. The Concept of Homonymy and Synonymy

It is worthwhile to study a little about the concept of Homonymy and Synonymy that these lexicographers used. The Synonymous lexica include both simple as well as compound words, whereas the homonymous lexica include only simple words and rarely the compounded words

The concept of Synonymy is not very well defined by these traditional lexicographers. As Ghatge(1973.28), points out, "The only criterion which they would like to use appears to be the idea of parivåttisahatva or the possibility of replacement. This is taken by them as a purely semantic criterion and a difference of gender alone is not sufficient to destroy the quality of paryäyatva We are thus left with their practice to see what they mean by a synonym " Words are classified into three categories, rüoha (Conventional), yaugika (derivational) and yogarüoha (derivational but restricted in usage by the convention). There are various kinds of relations which lead to

synonymy as listed by the traditional lexicographers. As far as the Homonymy is concerned there are several problems. The traditional lexicographers do not define the concept of anekārthaçabda At the same time the concept of Polysemy and Homophony remains undecided for these lexicographers.

IV. MODERN LEXICA OF SANSKRIT

The first Sanskrit Dictionary with western system of alphabetical order was the Sanskrit-English Dictionary compiled by Professor Horace Hayman Wilson and published in 1813. Two Indian works, viz. the Sabdakalpadruma compiled by Pandit Sir Raja Radhakanta Dev and the Vacasptya compiled by Pandit Taranatha Tarkavacaspati, followed suit.

A. Wörterbuch

But it was the Sanskrit-German Dictionary, Sanskrit Wörterbuch, compiled by Otto Böthlingk and Rudolph Roth, published from St. Petesberg during a span of 24 years-1852 to1875 - which is considered to be the benchmark as far as the lexicography related to Sanskrit is considered. The arrangement of meanings on historical principles is the methodology followed by this lexicon, which we see is adhered to by all later lexicographers. It drew its material from 450 books.

B. Monier Williams

This dictionary was compiled by Monier-Williams. Though he acknowledged his indebtedness to the Sanskrit-Wörterbuch, he worked for his dictionary on a plan of his own. It contained several features which had not been found in the Wörterbuch One can have an idea of those features from the subtitle of the dictionary, 'etymologically and philologically arranged, with special reference to cognate Indo-European languages.' The first edition of the dictionary was published in 1872 and the author, as soon as he became aware of the likelihood of his volume becoming out of print, set about preparations for a new improved and enlarged edition. He revised his original work in view of the criticisms, which the first edition invited from competent scholars and many Sanskrit texts that had been coming to light since the publication of the first edition. The second edition contains words and derivatives culled from more than four hundred Sanskrit texts. The author, while recording different shades of meaning of the words, has taken the meanings given by Indian lexicographers Kosa-karas and also the contexts of actual places of occurrence in the texts into consideration. The developments of meanings have been shown in chronological order. The author has also utilized the results of the research in comparative philology as available in his time. It drew its material from 515 books

The main feature of these two lexica is that they record the lexemes in the forms of stems. They also provide ample evidence in the form of the references where a meaning of a particular word occurred. 'kṅ' occurs as a conjunct consonant at the end of 'k'. Negative compound is treated as independent lexeme. Compounds with prepositions are mentioned under that preposition. Thus the words upakāra, upakåt, etc. are found under upa. The immediate constituent cut is the essential feature of these two lexica These are the lexica of both nominal as verbal stems. The derived roots are mentioned under the original roots. Thus the desiderative roots like cikerña are mentioned under kå However, not all such derived roots are treated in these lexica

C. Encyclopadic Dictionary of Sanskrit on Historic Principles

This is an encyclopedic dictionary drawing its lexemes from 1500 works, thrice more than that of the Monier Williams It records lexemes from Vedas up to the end of 18th century. The basic aim and plan of this dictionary seems to be the following-

(i) To indicate the earliest and the latest (in case it has become obsolete) occurrence of the vocable in Sanskrit Literature
(ii) To indicate whether a word existed throughout the history of that language or was confined to certain periods
(iii) To show the provenance of the word, i.e. whether it is used in many branches of learning or if it is system specific.
(iv) To show all available meanings, common as well as technical, and by arrangement, to show their historical relationship if traceable

It is an important question as to how these lexica are related to Sanskrit grammar and more specifically to the grammar of Panini? How does Panini treat them for different grammatical processes? But this will not be dealt with in detail in the present paper.

V. GRAMMATICAL LEXICA

It is important to note that there are strictly system specific lexica in Sanskrit, significant among them are the lexica built for the purpose of grammar. The most important among them is the Lexicon of verbal roots or dhatupatha as is traditionally known. This is believed to have been composed by Panini himself and awaits attention for its scientific arrangement from the point of view of grammatical functioning. The lexemes are arranged in ten groups according to the morphological operation they get in some specific environment. These groups are called as gaèas. Lexemes in each gaèa are further classified according to the application of certain types of suffixation. Based on this, the lexemes are

called "..padin"s. (ātmanepadin, parasmaipadin and ubhayapadin). Each lexeme is enjoined with a marker to indicate certain types of suffixations, certain augments, certain modifications and other grammatical operations. The rules in the grammar of Panini are aware of these classifications and this arrangement They often refer to the sub-classification amongst these gaèas We can say that this lexicon is very much system specific and leads to generation of specific forms. The rules related to these lexica generate adjectives and can be used to denote anything which performs a particular action. This is supported partly by the theory recorded by Yaska that all nouns are derived from verbal roots

As far as the lexica mentioned above deal with the nouns, Panini's grammar reclassifies all the nouns on formal basis He classifies them on the basis of the final element of the word Thus the first classification is - words ending in vowels and those ending in consonants. We do not find this kind of classification in any of the lexica

Panini's grammar talks of Karaka relations and the generative aspect of this grammar is closely related to the Karaka theory. With the help of this theory words can be generated from a given lexeme which will indicate the relation to a particular action.

Below given is a table in which a sample number of forms are shown to be generated with the help of grammatical rules of Panini from the lexicon of verbal roots, which can be used for coining new terms for scientific and technological purposes.

Root	Suffix	Karaka	Form	Meaning	Rule	Comment
Kr	A	kartå	Kara	One who does	3 1 133	Ad
	Ta	karma	Kåta	One wh.ch is done	3 2 102	Ad
	Ana	karaëa	Karaëa	The .nstrument of do.ng	3 3 117	Ad
	Tya	karma	kårya/kårya kartavya karaëeya	What Shou.d be done	3 1 93	Ad

I have not taken here the addition of the preverbs which are known as upasargas. They will be helpful in generating further meaning variations, and in fact, some new meanings altogether. In this manner any verbal root taken from the verbal lexicon can generate with the help of a grammatical rule and a suffix n number of forms

VI. APPLICATION OF PANINI'S GRAMMAR FOR CREATING NEW LEXICA:

It is this feature of the grammar of Sanskrit as well as that of compounding that can be effectively used for creating new technical terminology for Indian languages I here take the 30 + lexica published by the Government of Maharashtra, as a case study for illustrating this point in brief. Ever since 1967, the Directorate of Languages, Government of Maharashtra started the activity of preparing the lexica of technical terminology for all fields of Science and Technology as well as human sciences. The underlying principles for this activity are elaborately stated in the introduction to the lexicon for technical terminologies for Linguistics. It is a 19 points decree for such an activity These points include Exhaustive reference, Relevant reference, Amenability to testing referential fit, amenability to calibration, etc

These technical terms are two types- "technical terms collected into a technical terminology" and "technical terms collected into a technical nomenclature". For creating technical terminology of both these types verbal morphology as well as compounding features of Sanskrit are seen very effectively used in these lexica. Thus for example in the Lexicon on Linguistics, for 'acronym' the term coined is 'ädyakñare çabda'. This is of the second type mentioned above. It is formulated by the process of compounding and taddhita formation (derivational process of an adjective from a noun by adding a suffix) We can show the process as follows-

Adya + akñara ädyäkñara
adyakñara + in ädyakñarin

This becomes the adjective of çabda, which together with this adjective can be considered to be the coined technical term for acronym. Thus in order to coin the terminology for magnet and related words all the methods mentioned above are used. The technical nomenclature of a magnet is grasped in the word formed by adding the suffix "aka" after the root "cumb".

Cumb
Cumb + aka cumbaka

By the process of compounding other words are added to the above word and as many words are coined

Vidut + cumbaka vidyutcumbaka (electromagnet)
Nala + äkåti + cumbaka nälakåti cumbaka (horseshoe magnet)
Cumbaka + eya cumbakeya (magnetic)

It is possible to add a taddhita suffix after the compound. It is also possible to compound after the taddhita suffix is added

Thus we have following terms by the abovementioned process-

Cumbakéya pravardhi (magnet.c amplifier) Cumbak eya ak ña (magnetic axis)	([cumbaka + éya] + [pra + {vrdh + in}])
Cumbakéya paripatha (magnetic circuit)	([cumbaka + éya] + [pari + patha])
Cumbakéya nati (magnetic dip)	([cumbaka + éya] + [nam + ti])
Cumbakéya kñetra (magnetic field)	
Cumbaka sûci (magnetic needle)	
Cumbakana (magnetization)	(cumbaka + ana)
Cumbakatva (magnetism)	(Cumbaka + tva)
Cumbakatvamiti (magnetometry)	([Cumbaka + tva] + [ma + ti])
Cumbakana bala (magnetizing force)	

In the same way, in almost all the lexica mentioned above, we find this principle used. In future as well we can make use of this feature of Sanskrit grammar in order to coin new terminologies for upcoming branches in the field of Science and Technology. We can also borrow certain grammatical elements from other languages and some elements from Sanskrit for this purpose. Thus in Linguistics, the terms "rūpima" is coined for Morpheme and "svanima" for phoneme. These are created out of the Sanskrit roots "rūp" and "svan" respectively and the English morpheme "eme". In this way, keeping in mind the 19 points mentioned above it is possible to borrow linguistic elements from different Indian languages and terms can be coined

APPENDIX

Some more Lexica (traditional as well as modern)

1. Anekārthadhvanimaijare of Mahakshapanaka (H Homonymous)
2 Anekārthasamgraha of Mankha (H)
3 Anekārthasamgraha of Hemacandra (H)
4 Abhidhanaratnamalà of Halayudha
5 Anekārthasamuccaya of Sasvata (H)
6 Trikāëoaçeña of Purushottamadeva
7 Nanarthasamgraha of Ajaya (H)
8 Medineof Medinikara
9 Viçvaprakàça of Mahesvara
10 Viçvalocana of Sridhara
11. Dhaturatnakara
12. Dictionary of Panini by S M Katre
13. Nyayakoça by Jhalakikar
14. Memàmsakosa by Kevalananda Sarasvati
15. Dictionary of Sankskrit Grammar by K.V.Abhyankar
16. Sanskrit Catalan Dictionary

BIBLIOGRAPHY

Primary Sources:

[1] Ekārthanāmamālā-dvyakñaranāmamālā of Saubhari, Edited by Ekanath D. Kulkarni, Deccan College Postgraduate and Research Institute, Poona, 1955

[2] Koçakalpataru (Fascicule I & II), Edited by Madhikar M Patkar and K. V. Krishnamoorthy Sarma, Deccan College Postgraduate and Research Institute, Poona, 1966

[3] Dharaëekoça of Dharanidasa Edited by Ekanath D. Kulkarni, Deccan College Postgraduate and Research Institute, Poona, 1968

[4] Nānārtharatnamālā of Irugapa Dandadhinatha Edited by Bellikoth Ramachandra Sharma, Deccan College Postgraduate and Research Institute, Poona, 1954

[5] Nānārthamaijare of Raghava, Edited by K. V. Krishnamoorthy Sarma, Deccan College Postgraduate and Research Institute, Poona, 1954

[6] Nāmamālikā of Bhoja, Edited by Ekanath D. Kulkarni and Vasudeo Damodar Gokhale, Deccan College Postgraduate and Research Institute, Poona, 1955

[7] Paryāyaçabdaratna of Dhananjayabhatta Edited by Ekanatha D. Kulkarni and M. C. Dixit, Deccan College Postgraduate and Research Institute, Poona, 1971.

[8] Pramanadiya-namamala of Makarandadasa (Part I), Edited by Ekanath D. Kulkarni, Deccan College Postgraduate and Research Institute, Poona, 1968

[9] Paramānandeyanāmamālā of Makarandadasa (Part II), Edited by Ekanath D. Kulkarni, Deccan College Postgraduate and Research Institute, Poona, 1971.

[10] çivakoça of Shivadatta Mishra, Edited by R. G. Harshe, Deccan College Postgraduate and Research Institute, Poona, 1952

[11] çāradeyākhyanāmamālā of Harsakirti Edited by Madhukar M Patkar, Deccan College Postgraduate and Research Institute, Poona, 1951

[12] Siddhaçabdārëava of Sahajakirti, Edited by M. G. Panase, Deccan College Postgraduate and Research Institute, Poona, 1965

[13] Bhautikaçāstra Paribhāñā Koça, Directorate of Languages, Government of Maharashtra, 1988

[14] Bhāñāvijiāna va Vāimayavidyā paribhāñā koça, Directorate of Languages, Government of Maharashtra, 2001

Secondary Sources:

[1] M M.Patkar, "History of Sanskrit Lexicography" Munshiram Manoharlal, New Delhi, 1981.

[2] Madhav Deshpande, ? "Review of *History of Sanskrit Lexicography by M.M. Patkar* ", Language, Vol. 59, No. 4, 1983, p 933

[3] A M Ghatge, R N Dandekar, M.A Menendale (edit), "Studies in Historical Sanskrit Lexicography", Deccan College Postgraduate and Research Institute, Poona, 1973.

[4] Claus Vogel, "Indian Lexicography", A History of Indian Literature, Vol V, Scientific and Technical Literature, Fascicule 4, viii, Wiesbaden Otto Harrassowitz, 1979 DM 48

[5] A.M Ghatge, "Introduction" the Encyclopaedic Dictionary of Sanskrit on Historic Principles, Deccan College Postgraduate and Research Institute, Poona, 1976-1978

Web links:

Nāmalingānuçāsana [**Amarakosa**], Kanda 1 Input by Avinash Sathaye (sohum@ms uky edu, Pramod $ SV Ganesan % (The text is to be used for personal studies)

www.sub uni-goettingen de ebene 1 fiindolo gretil 1 sanskr 6 sastra 2 lex amarklhu.htm - 65k

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.19 Issues in Developing Corpus for Malayalam from Web as Source

S. A. Shanavas, University of Kerala

Abstract— **This paper is about developing corpora for Malayalam language from the web as the source. It discusses the issues faced when attempting to prepare text corpus from the web. Creating corpora from web or web as corpus is a new and emerging area. Source for this corpus development is Malayalam language texts in the net. Earlier it was developed by using keyboard, OCR, etc. Now it is towards the web. The quantity of texts in the web is enormous. The two possibilities with respect to web corpus are the web as the source for making corpus and considering the web itself as the corpus. Here the reference is for the former one. The advantages with web as source are significant as it include the selection of texts from various domains, less spelling errors, etc. There are some serious issues to be considered in developing such corpora; the problems relate to encoding, spilt of words, space alignments, non-uniformity or non-standardisation of fonts, text in different formats, copyright issues, selection of texts without banner and advertisements, popup files, etc.**

I. INTRODUCTION

ANY collection of more than one text can be called a corpus, but the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition The following list describes the four main characteristics of the modern corpus:

- Sampling and representativeness
- Finite size
- Machine-readable form
- Updating possible

The concept of carrying out research on written or spoken texts is not restricted to corpus linguistics. Indeed, individual texts are often used for many kinds of literary and linguistic analysis - the stylistic analysis of a poem, or a conversation analysis of a TV talk show, etc. However, the notion of a *corpus* as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways

A large collection of texts for any type of research is said to be corpus and for centuries people have been collecting manuscripts, books and newspapers for analysis of a very laborious nature. Thankfully, as technological advances make the computerized storage and access of large quantities of information easier, so the construction and use of text corpora continue to increase, and the potential for research has widened considerably. A corpus can also be thought of as a collection of texts gathered according to particular principles for some particular purpose. A corpus is valuable because together its component texts allow statements to be made about language as a whole. Corpus linguistics provides the methodology to extract meaning from texts.

Corpus Linguistics (CL) has now emerged as one among the major areas in language technology and linguistic study. Computational Linguistics (CL), Corpus Lexicography (CL) and Corpus Linguistics (CL) are the three 'CL's important to the Modern Language Technologists. Once it was rejected by relativists and generative grammarians like Chomsky and others who argued that corpus would not be helpful for linguistic analysis. Realising the importance and uses of corpus in wide areas of corpus development has now become the focus of attention. Taking as its starting point the fact that language is not a mirror of reality but let us share what we know, believe and think about reality, it focuses on language as a social phenomenon, and makes visible the attitudes and beliefs expressed by the members of a discourse community. As the corpus construction plays an important role in language technology, researchers are building their own corpora, solving problems independently, and producing project specific systems, which cannot easily be re-used. The process of corpus building is a new research area, which lacks standardization and appropriate tools. Corpora, however, have changed the way in which linguists can look at a language

The importance of corpora to linguistic study can be appreciated from the definitions given to it. It is interesting to note that the first definition suggests that corpora can help in studying language, while the second goes further and implies that categorical statements can be made on the basis of a well-defined corpus. This is the most basic type of linguistic corpus annotation - the aim being to assign to each lexical unit in the text a code indicating its part of speech. Part-of-speech annotation is useful because it increases the specificity of data retrieval from corpora, and also forms an essential foundation for further forms of analysis (such as syntactic parsing and semantic field annotation). Part-of-speech annotation also allows us to distinguish between homographs

Language corpora contain large amount of textual information in machine-readable form of whole texts or large continuous genre of text extracts from different fields. The texts can be whole books, newspapers,

S A. Shanavas is with Technology and Resource Center for Malayalam Department of Linguistics University of Kerala Karavattom Thiruvananthapuram Kerala (e-mail. drsasha2002@yahoo com)

journals, speeches etc, or consist of extracts of varying length. A corpus may contain single texts in single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*).

Attempts of corpus creation are going on at different parts of the world and Corpus linguistics is emerging as a separate discipline. English being the language of Science and Technology takes a major share in the net text. English language also leads in the development of corpus of millions of words. The important source of creation of corpora may be from e-texts or the web. Some English corpora in use are British National Corpus (BNC), American National Corpus (ANC), European Corpus Initiative Multilingual Corpus I (ECI/MCI), EMILLE CIIL, LOB, Brown Corpus, etc. The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources designed to represent a wide cross-section of British English

The first corpus in Indian languages is the *Kolhapur Corpus of Indian English (KCIE)* developed by S.V Shastri (1988) and his colleagues at Shivaji University, Kolhapur. It contains approximately one million words of Indian English. The corpora contain large amount of texts in machine-readable form of whole texts or collection of different genre of text from different fields. The Central Institute of Indian Languages (CIIL, Mysore) has initiated the project on development of corpus for Indian languages in the late eighties. CIIL has developed corpora for different Indian languages. The corpus for each language is of 3 million lexical items which has been developed by different agencies. A corpus consisting of 2,350,000 words is made available for Malayalam. Any other major attempt is not reported so for the language. They are neither widely representative nor balanced, nor even large with database Also there has never been any attempt to make error free or enlarge that corpus with regular argumentation of new text samples from various text types and sources.

The Malayalam corpora developed so far are either by using keyboard or OCR method. Now it is the turn of web There are so many advantages with web as a source than the earlier time. Text quantity is immense; it gets representation from almost all domains, time required is less, less errors, and so on.

II. USES OF CORPORA

The study of Linguistic phenomena through large collections of machine-readable texts of corpora is known as Corpus Linguistics. It is the study and analysis of data obtained from a corpus. Corpora are multifunctional resources. Corpora have uses in Linguistics, Natural Language Technology and in Natural Language Processing. One can retrieve appropriate information from corpora to employ in various linguistic studies such as lexicography design, writing grammars, semantic study of words, pragmatic analysis of texts, sociolinguistic study, discourse analysis, etc. (Leech and Fligelstone 1992: 129) Teachers, students, and researchers in a variety of fields (e g. languages, business studies, law, medicine and engineering) use corpora for teaching materials, class room exercises, and making dissertations, etc Software developers, engineers and programmers use corpora to develop reference tools, natural language processing applications etc. Corpora allow researchers not only to count categories in traditional approaches to language but also to observe categories and phenomena that have not been noticed before

Corpus based approach makes it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand Corpus based studies attempt to categorize registers, dialects, styles or individual literary works in terms of their linguistic association patterns

In Lexical Studies: Empirical data has been used in lexicography long before the discipline of corpus linguistics was invented. Samuel Johnson, for example, illustrated his dictionary with examples from literature and in the 19th Century the Oxford Dictionary used citation slips to study and illustrate word usage. A linguist who has access to a corpus, or other (non-representative) collection of machine readable text can call up all the examples of a word or phrase from many millions of words of text in a few seconds Dictionaries can be produced and revised much more quickly than before, thus providing up-to-date information about language. Also, definitions can be more complete and precise since a larger number of natural examples are examined

III. CORPORA AND GRAMMAR

Grammatical (or syntactic) studies have, along with lexical studies, been the most frequent types of research which have used corpora. Corpora make a useful tool for syntactical research because of

- The potential for the representative quantification of a whole language variety
- Their role as empirical data for the testing of hypotheses derived from grammatical theory.

Many smaller-scale studies of grammar using corpora have included quantitative data analysis (Schmied's 1993). There is now a greater interest in the more systematic study of grammatical frequency - for

example, Oostdijk and de Haan (1994a) are aiming to analyse the frequency of the various English clause types.

Since the 1950s the rational-theory based empiricist-descriptive division in linguistics has often meant that these two approaches have been viewed as separate and in competition with each other. However, there is a group of researchers who have used corpora in order to *test* essentially rationalist grammatical theory, rather than use it for pure description or the inductive generation of theory. The formal grammar is first devised by reference to introspective techniques and to existing accounts of the grammar of the language. The grammar is then loaded into a computer parser and is run over a corpus to test how far it accounts for the data in the corpus. The grammar is then modified to take account of those analyses which it missed or got wrong

IV. CORPORA AND SEMANTICS

The main contribution that corpus linguistics has made to semantics is by helping to establish an approach to semantics which is objective, and takes account of indeterminacy and gradience. Mindt (1991) demonstrates how a corpus can be used in order to provide objective criteria for assigning meanings to linguistic terms. Mindt points out that frequently in semantics, meanings of terms are described by reference to the linguist's own intuitions - the rationalist approach that we mentioned in the section on Corpora and Grammar. Mindt argues that semantic distinctions are associated in texts with characteristic observable contexts - syntactic, morphological and prosodic - and by considering the environments of the linguistic entities an empirical objective indicator for a particular semantic distinction can be arrived at

Another role of corpora in semantics has been in establishing more firmly the notions of **fuzzy categories** and gradience. In theoretical linguistics categories are usually seen as being hard and fast - either an item belongs to a category or it does not. However, psychological work on categorization suggests that cognitive categories are not usually "hard and fast" but instead have fuzzy boundaries. So it is not so much a question of whether an item belongs to one category or the other, but how often it falls into one category as opposed to the other one. By looking empirically at natural language in corpora it is clear that this "fuzzy" model accounts better for the data. clear-cut boundaries do not exist; instead there are gradients of membership, which are connected with *frequency* of inclusion

V. CORPORA IN THE TEACHING OF LANGUAGES AND LINGUISTICS

Resources and practices in the teaching of languages and linguistics tend to reflect the division between the empirical and rationalist approaches. Many textbooks contain only invented examples and their descriptions are based upon intuition or second-hand accounts. Other books, however, are explicitly empirical and use examples and descriptions from corpora or other sources of real life language data

Corpus examples are important in language learning as they expose students to the kinds of sentences that they will encounter when using the language in real life situations. Students who are taught with traditional syntax textbooks, which contain sentences such as- "Steve puts his money in the bank" are often unable to analyse more complex sentences such as- "The government has welcomed a report by an Australian royal commission on the effects of Britain's atomic bomb testing programm in the Australian desert in the fifties and early sixties."

Apart from being a source of empirical teaching data, corpora can be used to look critically at existing language teaching materials. Kennedy (1987a, 1987b) has looked at ways of expressing quantification and frequency in ESL (English as a second language) textbooks. Holmes (1988) has examined ways of expressing doubt and certainty in ESL textbooks, while Mindt (1992) has looked at future time expressions in German textbooks of English These studies have similar methodologies - they analyse the relevant constructions or vocabularies, both in the sample text books and in Standard English corpora and then they compare their findings between the two sets. Most studies found that there were considerable differences between what textbooks are teaching and how native speakers actually use language as evidenced in the corpora. Some textbooks gloss over important aspects of usage, or foreground less frequent stylistic choices at the expense of more common ones. The general conclusion from these studies is that non-empirically based teaching materials can be misleading and that corpus studies should be used to inform the production of material so that the more common choices of usage are given more attention than those which are less common

Corpora have also been used in the teaching of linguistics. Kirk (1994) requires his students to base their projects on corpus data which they must analyse in the light of a model such as Brown and Levinson's politeness theory or Grice's co-operative principle. In taking this approach, Kirk is using corpora not only as a

way of teaching students about variation in English but also to introduce them to the main features of a corpus-based approach to linguistic analysis.

A further application of corpora in this field is their role in computer-assisted language learning. Recent work at Lancaster University has looked at the role of corpus-based computer software for teaching undergraduates the rudiments of grammatical analysis (McEnery and Wilson 1993). This software - Cytor - reads in an annotated corpus (either part-of-speech tagged or parsed) one sentence at a time, hides the annotation, and asks the student to annotate the sentence by him/herself. Students can call up help in the form of the list of tag mnemonics, a frequency lexicon or concordances of examples. McEnery, Baker and Wilson (1995) carried out an experiment over the course of a term to determine how effective Cytor was at teaching part-of-speech learning by comparing two groups of students - one who were taught with Cytor, and another, who were taught via traditional lecturer-based methods. In general the computer-taught students performed better than the human-taught students throughout the term.

VI. TEXT ENCODING AND ANNOTATION

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as annotation. An example for annotating a corpus is part-of-speech tagging, or POS-tagging, in which information about each word's part of speech (verb, noun, adjective, etc.) are added to the corpus in the form of tags. If corpora is said to be unannotated it appears in its existing raw state of plain text, whereas annotated corpora has been enhanced with various types of linguistic information. Unsurprisingly, the utility of the corpus is increased when it has been annotated, making it no longer a body of text where linguistic information is implicitly present, but one which may be considered a repository of linguistic information. The implicit information has been made explicit through the process of concrete annotation.

For example, the form "pooyi 'go (past)" contains the implicit part-of-speech information "past tense verb" but it is only retrieved in normal reading by recourse to our pre-existing knowledge of the grammar of Malayalam. However, in an annotated corpus the form "pooyi 'go' " might appear as " pooyi 'go VZ", with the code VZ indicating that it is a past tense (Z) form of a lexical verb (VV). Such annotation makes it quicker and easier to retrieve and analyse information about the language contained in the corpus.

VII. FORMATS OF ANNOTATION

Currently, there are no widely agreed standards of representing information in texts; in the past, many different approaches have been adopted, some are more lasting than others. One long-standing annotation practice is known as COCOA references. COCOA was an early computer program used for extracting indexes of words in context from machine readable texts. Its conventions were carried forward into several other programs, notably the OCP (Oxford Concordance Program). The Longman-Lancaster corpus and the Helsinki corpus have also used COCOA references. Very simply, a COCOA reference consists of a balanced set of angled brackets (< >) which contains two entities:

- A code which stands for a particular variable name
- A string or set of strings, which are the instantiations of that variable

For example, the code "A" could be used to refer to the variable "author" and the string would stand for the author's name. Thus COCOA references which indicate the author of a passage of text would look like the following:

<A CHARLES DICKENS>
<A WOLFGANG VON GOETHE>
<A HOMER>

COCOA references only represent an informal trend for encoding specific types of textual information, e.g. authors, dates, and titles. Current trends are moving towards more formalised international standards of encoding. The flagship of this current trend is the Text Encoding Initiative (TEI), a project sponsored by the Association for Computational Linguistics, the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. Its aim is to provide standardised implementations for machine-readable text interchange. The TEI uses a form of document markup known as SGML (Standard Generalised Markup Language). SGML has the following advantages:

- Clarity
- Simplicity
- Formally rigorous
- Already recognised as an international standard

In the TEI, each text (or document) consists of two parts - a header and the text itself. The header contains information such as the following:

- Author, title and date
- The edition or publisher used in creating the machine-readable text
- Information about the encoding practices adopted.

VIII. RECENT DEVELOPMENTS IN WEB AS CORPUS

Web or World Wide Web is the new resource identified for corpus collection. There are some major initiatives in this direction. WebCorp is the pioneer in this field. Creating corpora from web or web as corpus is a new and emerging area. Here the materials data are immense, easily accessible and up-to-date representation of all types of texts and selection from all domains. The web is immense, free and available by a mouse click, hundreds and millions of words of texts can be used for all manner of language research Language scientists and technologists are turning to web as source of language data, because it is the only language source for the type of texts one is interested. It can be stored from the web as source web as corpus.

WebCorp is a suite of tools which allows access to the World Wide Web as a corpus - a large collection of texts from which facts about the language can be extracted. The WebCorp interface is similar to the interfaces provided by standard search engines. You enter a word or phrase, choose options from the menus provided and then press the 'Submit' button. WebCorp works 'on top of' the search engine of your choice, taking the list of URLs returned by that search engine and extracting concordance lines from each of those pages - examples of your chosen word or phrase in context. All of the concordance lines are presented on a single results page, with links to the sites from which they came.

Despite the fact that a growing body of work has shown that the World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette, 2003), many fundamental issues about the viability and exploitation of the Web as a linguistic corpus are just starting to be tackled, ranging from Web frequency distributions and registers to efficient handling of massive data sets and copyright. Research on the Web as corpus is currently at a very exciting stage: increasing evidences point to the enormous potential of Internet as a source of linguistic data, but it is still far from a working, fully-fledged linguists' search engine. The issues are discussed so far and have to be discussed while developing web corpus. Describe web corpus collection projects or modules for one part of the processes (crawling, filtering, language-id, tokenising, lemmatising, POS-tagging, indexing), exploring characteristics of Web data, from a linguistics/NLP perspective use crawled Web data for NLP purposes

The web text is immense, free and available just by a mouse-click It contains hundreds of billions of words of text and can be used for all manner of linguistic research. Language scientists and technologists are increasingly turning to the web as a source of language data, because it is so big, because it is the only available

source for the type of language one is looking to and it is free and instantly available. Web is the source from where more language data at different texts can be collected. The source for the development of the corpus is primarily from the net and the e-text available in the net

IX. MALAYALAM CORPUS - FROM WEB

Malayalam is one of the important Indian languages, which have got more web publication than any other regional languages. Malayalam produced hundreds of websites not only on the language and the literature but also in Malayalam scripts itself. Earlier texts were prepared in different codes -ASCII, ISCII, etc. and with different text editors. In web different codes with dynamic font or in separate font with downloading facility were used. But today because of the Unicode system large number of texts are prepared in it and uploaded in its original script.

Almost all leading dailies in the language have their online editions and about a hundred journals or weeklies got publication through the web. It can also be noticed that the dailies go for four to five editions or updating in a day. Most of the dailies and weeklies keep their archives for quite some time so that a user can access without much difficulty. Kerala Govt. sites, portals, commercial portals, sites belonging to educational institutions, Govt. publications, blogs and personal files sites etc make the web source extensive Source for corpus making can also be the digital texts prepared by different agencies either to publish or for other purposes

The practical issues when attempting to prepare a corpus for Malayalam can be divided into two. One is related to text, encoding issues, font and code, text format, etc. i.e., technical issues and the second one is language specific, i.e. related to spelling, spacing, borrowed lexical items, etc.

The steps involved in building a corpus are- selection of texts, data entry, data validation and a set of tools for management and retrieval of data. The text can be collected from different sources such as from web (Malayalam sites), newspapers, journals, magazines, medical reports, autobiography, government reports, diaries, notices etc. While collecting texts of Malayalam from web, categorization of them is an important and necessary part of documentation. The varieties of language we use in different situations referred as registers. (e.g., Medicine, literature, astrology) are categorized into different levels and the documents copied from sites are systematically added to our corpus under the specified register. This saves time, data duplication, memory space etc. Data collection will also be easy. Finally it can be traced that all Malayalam documents are not incorporated from the web sites or from all particular domain

The primary hurdle is that different agencies are using different softwares and codes for developing e-texts. Non-standardization of fonts or non-uniformity in spelling system is another key issue. Perhaps one text itself may put see in different spelling system. The text from different sources show in different fonts and spelling systems is an important issue for a corpus developer. Those texts collected from different sources, which are in different fonts and spelling systems, need to be carefully handled. One remedy for this is that the texts with different fonts have to be converted to a single or uniform font in Unicode system.

Lammatisation or identification of lexical items from compounds and complex constructions is another vital issue to be carefully looked into. Parts of speech identification for noun and verb also require additional attention. Space between lexical items, as in that of languages like English, Hindi, etc. cannot always been applied for Malayalam text. Position of lexical items with grammatical function expects that of finite verb, is another significant issue.

Following can be listed as some key issues with respect to text corpus

1. Non standardization of spelling difference in spelling system
2. Font differences, including that of half consonants (chillu)
3. Non spacing between lexical items
4. Sandhi changes
5. Borrowed lexical items
6. Part of speech differences
7. Position class of lexical items in different context

The issues related to web are

1. Problems relates with display
2. Banners advertising
3. Popup screens
4 Images and text as images
5 Different formats (HTML, PDF, XML, etc.)
6 Encoding issues

The concern for the time being is a general monolingual text corpus of the language containing data from all subject domains, texts, genres, types, and walks of life. This corpus will be properly analysed to be used for word level tagging, word count, letter count, frequency count, spell checker development, etc.

Tagged corpora with both linguistic and extra-linguistic information are required for different application. Extra-linguistic information of text type, publication details, name, age and sex of author, domain and register of the texts along with the source information will be included in the corpus. Linguistic information like analytical marker, part of speech code, lexical category mark, grammatical category, sentence type, semantic code, anaphoric reference, etc. needed to be incorporated to get maximum use of the corpus

The documents are copied from web sites and stored in the server with meta data The meta data or extra linguistic data about the document is the whole information about that document such as source web site address, date, author, font, etc. So each document is tagged with additional information, which can be used for further linguistic research and for data retrieval

Here is a text taken from an Internet edition of a newspaper. Fig.1 is gibberish display of the text. If we go to some site and try to get the display of it, we get a rubbish display

Fig. 1

©O¡J°·4 dj£EJw

¨d¡Y°dj£EJq°¨T ©O¡a ·©d¸t ©O¡tµ AY ·d¥ta o«gl·h¡·4°h¿. dj£E Y°T·¢i J¡k«h°Yv ©O¡a ·©d¸t ©O¡tµi°« CTi® °« h°si® °¨h¿¡« cT'4¢¶°U®. ©Jjq·¢v h¡±Yh¿. j¡Q ·¨h¡¶¡¨J C·j« o«gl·w ltU¢µ°lj°·4Y¡i°« J¡X¡«. Jr°·s¸u hËj« cT´°¼ AK¢©kE ·¡ ¨hV¢´v dj£EJq°¨Ti°« f¢o¢co® o®J¥w ±d©lmc dj£EJq°¨Ti°« h×°« ©O¡a Tk¡o°Jw ©O¡J°·4Y°« ©Jo¡J°·4Y°« dY¢l°l¡t·Jq¡X¢©¸¡w. ©O¡a ·« ©O¡t·k°« AY¢¨Ê l¢k®dc1¢k¥¨T ©J¡T¢Jw oØ¡a¢ °·4Y°« luJ¢T h e¢1¡ ±dlt·ch¡1¢ h¡s¢1¢j¢ °·4°. ©Jjq·¢v Jr¢°ltn« cT·4 Fo® Fo®. Fv.o¢ dj£Ei°¨T ©O¡a ··w ©O t'4Y¢c°d¢·4¢k°« J¡n®±T£1 ¡J°« D 4Y D©a ·¡Lo®Zc®h¡J°« Alj°¨T CTc¢k ¡J°« Dw¨¸¶ lk¢¨1¡J° o«M«Y¨1 4 D³4Y¸1¡X® C©Y¸×¢ A©c§n¢ °'4 o¢ f¢

Same Malayalam text in a different font in language script (*MLW-TTKarthika*)

©ഛക്കഷഗ്ഗസ്വ റഷ്കക്ഷഗ്ഗ്ഖം
ച്ചറക്കഥഗ്ഗറഷ്കക്ഷഗ്ഗ്ഖുഗ്ഗച്ചം ©ഛക്കമൺ©റപ്പ് ©ഛക്കന്ദേ അഥൺറഗ്ഗഗ്ഗോ
ട്ടഴഹച്ചവക്കസ്വഗ്ഗവയ്യ. റഷ്കക്ഷഗ്ഗ ഥഗ്ഗംപ്പക്ഷ്ശ ക്കെസട്ടവഗ്ഗ്ഥ് ©ഛക്കമൺ©റപ്പ്
©ഛക്കന്ദേശഗ്ഗട്ട ഇംശണ്ണന്തഗ്ഗട്ട വഗ്ഗശൺണ്ണന്തഗ്ഗച്ചവയ്യക്കട്ട ര്ംസക്ക്¶ഗ്ഗസ്ഥണ്ണ ©ഖഷ്യ·ക്ക് വക്കദ്ദഥവയ്യ, ഷക്കന്ധൺച്ചവക്ക്¶ക്കച്ചഖ ഇ·ഷ്ടട്ട ട്ടഴഹവ്ലം ഹജ്ഞേക്കന്ദഗ്ഗഹഷ്ഗ്ഗസ്ഥഥക്കശഗ്ഗട്ട ഖക്കതക്കട്ട. ഖ്യഗ്ഗ·പ്പൗ വഹ്ലഷട്ട ര്ംന്തഗ്ഗസ്വ അഗക്ക്ഷ©സസ്സൺക്ക ച്ചവഡ്ഢക്കന്ത് റഷ്കക്ഷഗ്ഗഖ്യഗ്ഗച്ചംശഗ്ഗട്ട ഉക്ക്ഖിക്കറിണ്ണ ിണ്ണഖവ്ലം ദ്ദറ©ഹാര റഷ്കക്ഷഗ്ഗഖ്യഗ്ഗച്ചംശഗ്ഗട്ട വണ്മഗ്ഗട്ട ©ഛക്കമൺന്തംസക്കിഗ്ഗഖം ©ഛക്കക്ഷഗ്ഗസ്വഥഗ്ഗട്ട ©ഖീക്കഖഗ്ഗസ്വഥഗ്ഗട്ട റഥക്കഷഗ്ഗഹക്കേ·ഖുക്കതക്ക©പ്പക്കം. ©ഛക്കമൺട്ട ©ഛക്കേ·സഗ്ഗട്ട അഥക്കച്ചസ്ല ഹക്കസണ്ണറശക്കസഗ്ഗച്ചം ©ഖക്കംക്കഖം ിസ്ഥക്കമക്കന്തഗ്ഗസ്വഥഗ്ഗട്ട ഹൗഖക്കം വക്കലക്കശക്ക ദ്ദറഹ·രവക്കശക്ക വക്കക്കൈശക്കഷക്കന്തഗ്ഗസ്വഗ്ഗ. ©ഖഷ്യ·ക്ക് ഖ്യക്കസ്വഹട്ടേ ര്ംസ എിണ്ണ. എിണ്ണ. വ്ിക്ക റഷ്കക്ഷഗ്ഗശഗ്ഗച്ചം ©ഛക്കമൺന്ലം ©ഛക്കസ്വേഥക്കശഗ്ഗറഗ്ഗക്കസക്കസഗ്ഗട്ട ഷക്കിണ്ണദ്ദംക്ഷശ ന്തക്കഷഗ്ഗട്ട ഉസ്വഥ ഉ©മൺക്കറ്റിണ്ണദരണ്ണവക്കഷഗ്ഗട്ട അഹഷഗ്ഗച്ചം ഇംര്ക്കസന്തക്കഷഗ്ഗട്ട ഇംച്ചപ്പ്¶ ഹസക്കച്ചശക്കഷഗ്ഗ ട്ടടട്ടഥച്ചസ്വ ഉമ്മഥക്കശക്കതണ്ണ ഇ©ഥപ്പണക്ക അ©രങ്ങിക്കന്തഗ്ഗസ്വ ിക്ക ഉക്ക ച്ചഎി. ഖ്യക്കസ്വ മക്കഹിട്ട ഉ്യഗ്ഗവക്കരച്ചപ്പ്¶ ച്ചച്ചു©ന്തക്കംഥക്കച്ചശ ©ഉക്കയക്കപ്പക്കന്ദഥണ്ണ ച്ചറക്കഥഗ്ഗറഷ്കക്ഷഗ്ഗക്ക ിദ്ദസ്ഥമക്കശച്ച·

CIIL corpus is now available on net for downloading It is in PDF format.

<5sch04><Aesthetics><Text Book
(School)><1988><Book><കേരള പാഠാവലി
സ്റ്റാന്റേഡേഡ് 10><****><15>

1 വേലയിൽ വിളയുന്ന വിദ്യാഭ്യാസം
(മഹാത്മാഗാന്ധി 1869 1948, രാഷ്ട്രപിതാവ്
ആദർശങ്ങൾ കർ ്മഫലത്തിലെത്തിച്ച മഹാൻ.
ജീവിതത്തിന്റെ എല്ലാ മേഖലകളെയും സമഗ്രമായി
വീക്ഷിക്കുന്ന
'എന്റെ സത്യാന്വേഷണ പരീക്ഷണങ്ങൾ'
അദ്ദേഹത്തിന്റെ
ആത്മകഥയാണ് ഗാന്ധിജിയുടെ ചിന്തകളും
ദർശനങ്ങളും
സമാഹരിച്ച് മലയാളത്തിൽ ഏഴു ഭാഗങ്ങളായി
'ഗാന്ധിസാഹിത്യം ' എന്നപേരിൽ ഒരു
ഗ്രന്ഥപരമ്പര
പ്രസിദ്ധീകരിച്ചിട്ടുണ്ട് ഒരു
നവസമൂഹനിർമ്മിതിഗുവേണ്ടചി

X. WEB SOURCE FOR MALAYALAM CORPUS

Thousands of pages are there in Malayalam in the web with Malayalam scripts. The online dailies texts from journals, e-texts, etc. enable us to closely observe the day-to-day variation of language. The advantages of online texts are that they have less spelling errors, true representation from all domains, etc.

Preference will be given to projects where Web data are downloaded and processed directly, rather than via search engine interfaces. The required data can be collected from the Net or e-texts, may be by manual selection and store or automatically through a search

program to store. Malayalam works are being developed and printed in abroad also. Thousands of literary and other works published and printed regularly have a literature of world standards. Malayalam produces more than any other regional language in the country. Cyber Malayalam is more than in print, and that include text types like books, newspapers, journals, blogs, portals, etc

A model of data collected and saved in *.doc format.

<Site address: http www.malayalamresource centre.org/Mrc/Karshikam rubber rubber1.html, Root. Kaarshikam rubber, Author: C-DAC, Date: 03 05 07, Remark:
The agriculture of Kerala describing about rubber, cardamom, cashew nut, clove, coconut, paddy, tea, and pepper.>

ഭാരതത്തിലെ റബ്ബർതോട്ടവ്യവസായം ശതാബ്ദിയിലെത്തി നില്ക്കുകയാണിന്ന് പത്തുലക്ഷത്തോളം വരുന്ന റബ്ബർ കർഷകരിൽ എൺപത്തേഴുശതമാനത്തോളം ചെറുകിട കർഷകർ-ഭാരതത്തിലെ റബ്ബർ കൃഷി മേഖലയുടെ മാത്രം പ്രത്യേകതയാണിത് ആഗോള വല്ക്കരണത്തിന്റെ ഭാഗമായി സംജാതമാകുന്ന പ രിതസ്ഥിതികൾ തരണം ചെയ്യുന്നതിനും എന്നെന്നും റബ്ബർതോട്ട വ്യവസായത്തെ സാധാരണ കർഷകന്റെ ആശാകേന്ദ്രമായി നില നിർത്തുന്നതിനും റബ്ബർകൃഷിയുടെ ശാസ്ത്രീയത പ്രയോഗിക തലത്തിൽ എത്തിച്ചേ മതി യാകൂ ഉല്പാദനചെലവു ചുരുക്കിയും ഉല്പാദനക്ഷമത വർദ്ധിപ്പിച്ചും ഉല്പന്നത്തിന്റെ ഗുണ മേന്മ ഉറപ്പുവരുത്തിയും മാത്രമേ കൃഷിയിടങ്ങളെ ആദായകരമാക്കാൻ സാദ്ധ്യമാകുകയുള്ളു ഭാരതത്തിലെ റബ്ബർ കൃഷി ഇന്ന് ഒരു പരീക്ഷണഘട്ടത്തിലെത്തിയിരിക്കുകയാണ് മറ്റു നാണ്യവിളകളുമായി തട്ടിച്ചുനോക്കിയാൽ ഉല്പാദനക്ഷമതയിലും ആദായത്തിലും ഉന്നതമായ സ്ഥാനം റബ്ബർകൃഷിക്കുണ്ടെങ്കിലും ഉല്പാദനഗുണനിലവാരവർദ്ധനവിലൂടെ വരുമാനമുയർത്താൻ ഇനിയും സാധ്യത ധാരാളമുള്ള കൃഷിയാണ് റബ്ബർ പ്രധാന റബ്ബർ ഉല്പാദക രാഷ്ട്രങ്ങളുമായി തട്ടിച്ചുനോക്കിയാൽ ഉല്പാദനക്ഷമതയിൽ നാം ഒന്നാം സ്ഥാനത്താണ് ഭാരതത്തിൽ റബ്ബറിന്റെ ഉല്പാദനക്ഷമത 1567 കിലോഗ്രാമാണ് എന്നാൽ 2500 കിലോഗ്രാംവരെ ഉല്പാദനക്ഷമതയുള്ള തോട്ടങ്ങൾ ഭാരതത്തിൽ ധാരാളമുണ്ടെന്നതാണ് വസ്തുത ഈ വസ്തുത കണക്കിലെടുത്താൽ ഇന്ത്യയിലെ ഉല്പാദനക്ഷമത 1567 കിലോഗ്രാമിൽ നിന്നും 2500 - 2000 കിലോഗ്രാമായി വർദ്ധിപ്പിക്കാൻ സാധിക്കാമെന്നത് നിസ്തർക്കമായ സംഗതിയാണ്.

(MLW-TTKarthika)

XI. ISSUES IN THE DESIGN PHASE OF MALAYALAM CORPUS

First stage in Malayalam corpus building is the collection of data from more than one source or from one genre. So the corpus builder should have a clear view about the categories and subcategories of the texts, which are to be collected

- Non standard / different fonts, Text in different format (HTML, PDF, etc.)
- Banners, popup screens, advertisements
- Text as images
- Spelling errors
- Alignment issues

Classification of texts belonging to different domains is one of the main issues related to the classification of texts. Subsections of texts classified by date, subject-matter, region, age-group, sex, type, style, register, etc. are also to be considered seriously.

Some of the Malayalam sites that can be used for data collection:

Online newspapers

1. www keralakaumudi com
2. http: www.manoramaonline.com cgi-bin/MMOnline dll portal ep/home do?newUser yes
3. http: www.thejasonline com java-thejason index jsp
4. http www.deshabhimani com
5. http: www deepika.com
6. http www keralaonlive com
7. http: www.mathrubhumi.com
8. http: www.mangalam com
9. http: sify com malayalam

Periodicals

1. http: www.vellinakshatram.com 934 index.asp
2. http: www vellinakshatram com/kala 1655 index.asp
3. http: www.vellinakshatram.com fire 168 index.asp
4. http: www.manoramaonline.com cgi-bin MMOnline.dll portal ep/home.do?tabId 5
5. http://www.jaalakam.com/
6. http: www.weblokam.com
7. http: www.malayalamvarikha.com
8. http: www.vellinakshatram.com muhurtham 57 index.asp
9. http: www.vellinakshatram.com ayurarogyam 58 index.asp
10. http: www.vellinakshatram.com snehitha 77 index.asp

Other useful sites

1. http: www.tapioca.in (Malayalam search engine)
2. http: www.puzha.com
3. http: www.chintha.com malayalam/blogroll php

XII. CONCLUSION

A corpus is not simply collection of texts; rather a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus, therefore, depends upon what it is meant to represent. A corpus can be developed from web as a source. There are a lot of advantages in web as a source. Similarly a number of issues are to be solved for creating corpus from the web

Some of the notable advantages of web as a source for corpus development are:

- New and emerging area
- Text data is immense, free and available by the click of a mouse
- Easily accessible
- Up-to-date representation
- Selection from all domains
- Can be stored directly from the web with or without modification
- Web can also be treated as corpus
- The web text in different formats need to be unified
- Hundreds and millions of words, texts
- Language scientists and technologists are more interested
- Because it is the only language source for the type of texts one is interested
- Access to web corpus is ubiquitous and size exceeding all previous corpuses.

The major issues that are creating hurdles for creating corpus from the web are:

1. Non standardization of spelling – difference in spelling system
2. Font differences, including that of half consonants (chillu)
3. Non spacing between lexical items
4. Sandhi changes
5. Borrowed lexical items
6. Part of speech differences
7. Position class of lexical items in different context

The issues related to web are

1. Problems relates with display
2. Banners advertising
3. Popup screens
4. Images and text as images
5. Different formats (HTML, PDF, XML, etc.)
6. Encoding issues

Malayalam corpus can be developed from the web as a basic resource for language analysis and research. Some of the above issues can be solved easily and some others require much research and analytical study. This is high time for us to turn our attention towards web as corpus or web as source for corpus, and of course, web is the real resource or mine for such development of our linguistic research and description.

REFERENCES

[1] Biber, D., S. Conrad and R. Reppen 1998. "Corpus Linguistics: Investigating Language Structure and Use" Cambridge: Cambridge University Press

[2] Bharati Akshar, Chaitanya Vineet, Sangal Rajeev 1996 Natural Language Processing- A Paninian Perspective, New Delhi Printice - Hall

[3] Dash, N.S., and B B Chaudhuri, 2000. "The process of designing a multidisciplinary monolingual sample corpus" International Journal of Corpus Linguistics. 5(2): 179-197.

[4] Dash, N. S., "Language Corpora: Present Indian Need" Indian Linguistics

[5] Dash, N.S "Utilization of Corpora in compilation of a monolingual general dictionary in electronic form"

[6] Kilgarriff, Adam "Web as Corpus" Computational Linguistics Vol. V No. N

[7] Orasan Constantin, Ramesh Krishnamurthy, UK "An Open Architecture for the Construction and Administration of Corpora"

[8] Ruslan Mitkove, The Oxford Handbook of Computational Linguistics, Oxford: Oxford Press, 2001

[9] Report of the Committee on Malayalam Character Encoding and Keyboard Layout Standerdisation, Govt. of Kerala, 18th December 2001

[10] Volk Martin "Using the Web as Corpus for Linguistic Research", article from web

[11] www.webcorp.org

[12] www.kilgarriff.co.uk

This paper was presented at LRIL-2007 National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.20 Handling of Case Markers for Designing UNL Based Punjabi Language Server

Parteek Bhatia, Thapar Institute of Engineering and Technology, Patiala, Punjab

Abstract **Case marker plays an important role in designing a Language Server. In UNL based Machine Translation system Language servers are used to convert the source language text to UNL with the help of Enconverter and Deconverter is used to convert the UNL text to target language. In order to implement the complete multilingual machine translation on Internet, it requires the language server for each language which encompasses Enconverter and Deconverter for that language. In this paper we discuss the handling of case markers for designing Punjabi Language Server. It can play an important role in designing and developing Enconverter and Deconverter for Punjabi Language.**

I. INTRODUCTION

The World Wide Web represents a formidable tool for communication and information access. With simple equipment, it is possible to access innumerable documents about a huge variety of topics, from any place around the world. However, despite the abundance of information, languages very often cause problems. [1] Most of the web pages today are written in few most commonly used languages like English, French, Chinese, etc.; it becomes difficult for a person with insufficient knowledge of these languages to access and use this tool of communication and information. This has prompted the need to devise means of automatically converting the information from one natural language to another natural language, called Machine Translation. [3] This process needs syntactic and semantic analysis of both source and target languages.

In case of Deconverter there are different phases for the generation of meaningful Punjabi sentences from the UNL The process of deconversion involves syntax planning, case marker generation, and morphology phase. [2] After the syntax-planning phase, which is aimed at generation of proper sequence of words, case marking phase initiates. Case Marker phase used to express the complete contents of the sentence. In the deconverter process the first task is to parse the UNL file. The nodes are generated in the target language with the help of L-UW dictionary. After generation of target language nodes they are ordered in Syntax planning phase according to the grammatical details of target language. Then Case marker phase fills the nodes generated from syntax planning phase with the appropriate case marker for the target language depending upon the use of the relations in the UNL text. For this purpose it takes into consideration Relational Morphology of the target language

II. UNIVERSAL NETWORKING LANGUAGE

Universal Networking Language (UNL), developed at UNU, is a formal language for representing the meaning of natural language sentences The motivation behind UNL is to develop an interlingual representation such that semantically equivalent sentences of all languages have the same interlingual representation. Information expressed in UNL can be converted into the native user's native language with higher quality and fewer mistakes than the computer translation systems. In addition UNL unlike natural language is free from ambiguities.

The UNL represents information sentence by sentence. Each sentence is converted into a directed hyper graph having concepts as nodes and relations as arcs

The knowledge within document is expressed in three dimensions: [4]

- Word knowledge is expressed by Universal Words (Uws).
- Relating UWs through a set of UNL relations capture concept Knowledge
- Speakers view, aspect, time of event, etc. are captured by UNL attributes.

Example UNL

Original sentence - John breaks the rules

UNL Corresponding to sentence is as follows:

agt(break(icl>do) @entry, 'John")
obj(break(icl>do), @entry, rule @generic @pl)

In this, 'agt' and 'obj' are the relations where words starting with character "@" are attributes corresponding to each universal words, i.e. 'Break' is the universal word in this sentence and words like @pl, @generic are the attributes describing the universal word

III. CASE MARKERS

Case is a category of morpho-syntactic properties which distinguishes the various relations that a noun phrase may bear to a governing head. Some of these relations are purely syntactic in nature. In the Indian linguistic system - descended from Sanskrit - the case constructs are called kaaraks. [5] As in the traditional understanding, they denote the relationship of the nominal with the main verb of the clause. The case structure in Punjabi is complex. An exhaustive study of the kaarak system with view to analyzing Punjabi into UNL has been carried out

Case marker phase apply proper case marker for each and every relation in the given UNL expression.

Parteek Bhatia is with Computer Science & Engineering Department, Thapar Institute of Engineering and Technology Patiala Punjab (e-mail: parteek.bhatia@tiet.ac.in)

There are total forty-five relations defined in UNL specifications and for each relation different case markers are used depending upon the grammatical details of that language

Here we discuss the all forty-five relations defined in UNL [4] specifications with corresponding case markers for Punjabi language, which are used to design Case marker data file for Punjabi Language Server. Case marker for each relation in UNL is shown in table 1.

A Case Marker data file contains one or more set of constraints for each relation and each of these sets map to different case markers. So given a node with all its attributes including lexical attributes from dictionary, we search the database for appropriate rule which the node satisfies and accordingly the case markers are initialized for the case markers

Table 1 : Case Markers for UNL relations

No	UNL Relation	Description	Punjabi Case Marker	Example
1	agt	Agent i.e. a thing which initiates an action	ਨੇ	ਰਾਮ ਨੇ ਚੱਲ ਖਾਧੇ
2	and	conjunction i.e a conjunctive relation between concepts	ਤ	ਰਾਮ ਤ ਸਾਮ ਦੱਸਤ ਹਨ
3	bas	basis i.e. indicates a thing used as the basis (standard) of comparison	ਤੋਂ	ਸੱਤ ਦਸ ਤੋ ਛੱਟਾ ਹੈ
4	ben	beneficiary i.e indicates an indirectly related beneficiary or victim of an event or state	ਦੇ ਲਈ	ਦੇਸ਼ ਦੇ ਲਈ ਮਾਨ ਦੇਣਾ
5	cag	co-agent i.e. indicates a thing not in focus that initiates an implicit event that is done in parallel	ਦੇ ਨਾਲ	ਰਾਮ ਸਾਮ ਦੇ ਨਾਲ ਗਿਆ
6	cao	co-thing with attribute i.e indicates a thing not in focus that is in a parallel state	ਦੇ ਨਾਲ	ਮਾ ਧੀਆ ਦਆਵਾਂ ਦੇ ਨਾਲ ਰਾਮ ਸੁਰਖਿਅਤ ਹੈ
7	cnt	content i.e. indicates the content of a concept	null	ਪੈਸੇ ਖਣੇ ਮਾਣ ਦਾ ਖਤਰਾ
8	cob	affected co-thing	ਦੇ ਨਾਲ	ਰਾਮ ਮਿੱਤਰਾ ਦੇ ਨਾਲ ਦਿੱਲੀ ਗਿਆ
9	con	condition i.e. indicates a non-focused event or state that conditions a focused event or state	ਜੇ	ਜੇ ਤੁਸੀ ਥੱਕ ਗਏ ਹੋ ਤਾ ਅਸੀ ਘਰ ਚੱਲਦ ਹਾਂ
10	coo	effected co-thing I.e. indicates a co-occurrent event or state for a focused event or state	ਉਦੋ,ਜਦ	ਜਦੋ ਮੈ ਵਾਪਿਸ ਆਵਾ ਉਦੋ ਤੁਸੀ ਮਾਣਾ ਹੈ
11	dur	duration i.e indicates a period of time during which an event occurs or a state exists	ਦੇ ਵੇਲੇ	ਸਵੇਰ ਦੇ ਵੇਲੇ ਮੌਸਮ ਚੰਗਾ ਹੰਦਾ ਹੈ
12	equ	effected co-thing i.e. indicates an equivalent concept	null	ਕਲਮ ਇੱਕ ਲਿਖਨੇ ਕੀ ਚੀਜ
13	fmt	range/from-to i.e. indicates a range between two things	ਤੋ	ਪਟਿਆਲਾ ਤੋ ਅਪਨ ਤਕ
14	frm	origin i.e indicates an initial state of a thing or a thing initially associated with the focused thing	ਤੋ	ਪਟਿਆਲਾ ਤ ਦਿਲੀ ਦੂਰ ਹੈ
15	gol	goal state i.e. indicates a final state of object or a thing finally associated with the object of an event	ਵਿਚ	ਰਾਮ ਪਟਿਆਲਾ ਵਿਚ ਰਹਿੰਦਾ ਹੈ
16	icl	indicates an upper concept or a more general concept	ਤਰਾ ਦਾ	ਕੱਤਾ ਇਕ ਤਰਾਂ ਦਾ ਜਾਨਵਰ ਹੈ
17	ins	Instrument i.e. indicates an instrument to carry out an event	ਦੇ ਨਾਲ	ਕਲਮ ਦੇ ਨਾਲ ਲਿਖ

18	int	Intersection i.e. indicates all common instances to have with a partner concept	null	ਇਕ ਕਿਤਾਬ
19	iof	an instance of i.e. indicates a class concept that an instance belongs to	null	ਪਟਿਆਲਾ ਪੰਜਾਬ ਦਾ ਸਹਿਰ ਹੈ
20	man	manner i.e. indicates a way to carry out an event or the characteristic s of a state	null	ਮਲਦੀ ਚਲੋ
21	met	method or means i.e indicates a means to carry out an event	ਦੇ ਨਾਲ	ਪੱਥਰ ਹਥੌੜੇ ਦੇ ਨਾਲ ਤੋੜਿਆ ਗਿਆ
22	mod	modification i.e. indicates a thing that restricts a focused thing	ਦਾ	ਚੰਦਰਮਾ ਧਰਤੀ ਦਾ ਉਪਗ੍ਰਹਿ ਹੈ
23	nam	name i.e. indicates a name of a thing	null	ਓਹ ਰਾਮ ਹੈ
24	obj	affected thing i.e. indicates a thing in focus that is directly affected by an event or state	ਤੇ	ਰਾਮ ਨੇ ਸੀਤਾ ਤੋ ਕੰਮ ਕਰਾਇਆ
25	opl	affected place i.e. indicates a place in focus affected by an event	ਤ	ਵਿਚਕਾਰ ਤ ਕੱਟੋ
26	or	disjunction i.e. indicates a partner to have disjunctive relation to	ਜਾ	ਅਸੀ ਅੱਜ ਇਥੇ ਰਹਗੇ ਜਾ ਨਹੀ
27	per	proportion rat e distribution i.e. indicates a basis or unit of proportion, rate or distribution	ਵਾਰ	ਮੈ ਚਾਰ ਵਾਰ ਉਥੇ ਗਿਆ

28	plc	indicates a place where an event occurs, or a state that is true, or a thing that exists	ਵਿਚ	ਹਸੋਇ ਵਿੱਚ ਖਾਨਾ ਬਨ ਰਿਹਾ ਹੈ
29	plf	initial place i.e. indicates a place where an event begins or a state that becomes true	ਤੋ	ਪਟਿਆਲਾ ਤੋ ਮਪਾਨ ਤਕ
30	plt	final place i.e. indicates a place where an event ends or a state that becomes false	ਤਕ	ਪਟਿਆਲਾ ਤਕ ਸਫਰ
31	pof	part of .e. indicate a concept of which a focused thing is a part	ਦਾ/ਦੀ/ਦੇ	ਕਿਤਾਬ ਰਾਮ ਦੀ ਹੈ
32	po s	possessor .e. indicates the possessor of a thing	null	ਮਰੀ ਕਿਤਾਬ
33	ptn	partner .e. indicates an indispensable non focused initiator of an action	ਦੇ ਨਾਲ	ਰਾਮ ਦੇ ਨਾਲ ਸ਼ਾਮ ਰਹਿਦਾ ਹੈ
34	pur	purpose i.e. indicates the purpose or objective of an agent of an event or the purpose of a thing that exists	ਦੇ ਲਈ	ਰਾਮ ਨੰ ਬਚਾਉਣ ਦ ਲਈ ਸ਼ਾਮ ਝਠ ਬਲ ਰਿਹਾ ਹੈ
35	qua	quantity .e. indicates the quantity of a thing or unit	Null	ਦੋ ਚਹ
36	rsn	reason .e. indicates a reason why an event or a state happens	ਦੇ ਕਰਨ	ਮੀਹ ਦ ਕਾਰਨ ਸਾਰੀ ਫਸਲ ਖਰਾਬ ਹ ਗਈ

37	scn	scene i.e. indicates a scene where an event occurs, or state is true, or a thing exists	ਵਿਚ	ਰਾਮ ਦੀ ਤਸਵੀਰ ਕਲ ਟੀ ਵੀ ਵਿਚ ਦਿਖਾਈ ਗਈ
38	seq	sequence i.e. Indicates a prior event or state of a focused event or state	ਦੇ ਬਾਅਦ	ਤੀਨ ਦੋ ਬਾਅਦ ਚਾਰ
39	src	indicates the initial state of an object or thing initially associated with the object of an event	ਤੋਂ	ਬਤੀ ਲਾਲ ਤ ਪਰੀ ਹਈ
40	tim	time i.e indicates the time an event occurs or a state is true	ਨੂੰ	ਮੰਗਲਵਾਰ ਨੂੰ ਚਲਾਂਗੇ
41	tmf	initial time i.e indicates the time an event starts or a state becomes true	ਤੋਂ	ਸਵੇਰ ਤੋ ਲੈਕੇ ਸਾਮ ਤਕ
42	tmt	final time i.e indicates a time an event ends or a state becomes false	ਤਕ	ਸਾਮ ਤਕ ਘਰ ਪਹੰਚਾਗੇ
43	to	indicates a final state of a thing or a final thing (destination) associated with the focused thing	ਦੇ ਲਈ	ਗਡੀ ਲੰਦਨ ਦ ਲਈ ਹੈ
44	via	an intermediate place or state	ਵਲੋ	ਪਟਿਆਲਾ ਵਲੋ ਜਲਦੀ ਪਹੰਚਾਗੇ
45	aoj	A thing which is in a state or has an attribute	Null	ਲਾਲ ਕਲਮ

IV. RESULTS

In the following examples use of case markers are shown. From these examples we can understand how much case markers are important in natural language generation.

Example 1:
(Original sentence)

Ram ate dinner.

{unl}
agt (eat.@past.@entry, Ram (agt<person))
obj (eat.@past.@entry, dinner (icl<food))
{unl}

Output without case marker-

ਰਾਮ ਖਾਣਾ ਖਾਧਾ

Now this sentence is not meaningful. To generate meaningful sentences we have to apply the case markers within the sentences. If '***agt***' has parent UW with *@past* attribute and it is transitive verb, and if child UW is noun, then put case marker ਨੇ after the child UW, i.e. Ram.

So ultimately the output after applying the case marker rule will be-

ਰਾਮ ਨੇ ਖਾਣਾ ਖਾਧਾ

Example 2:
(Original sentence)

Pest comes out from egg

(Partial UNL)
src(come out(icl>happen). @custom.@entry,
egg(ic.>foodstuff).@def)

Output without case marker rule-

ਬਚਾ ਅੰਡੇ ਬਾਹਰ ਆਉਂਦਾ ਹੈ

Again we can see that this sentence is meaning, so to generate a meaningful sentence case marker rule will be applied. It says that if child UW is noun and relation is '**src**' (source) then put next case marker of child as ਤੋਂ .

Ultimately after applying the case marker rule output is:

ਬਚਾ ਅੰਡੇ ਤੋਂ ਬਾਹਰ ਆਉਂਦਾ ਹੈ

V. CONCLUSION

In this paper we discussed the case markers for the Punjabi language server based on UNL relations. In this we came to know how case markers are important for any language generation. We showed this with the help of examples. For any language, study of case constructs are very important for the generation of meaningful sentences

REFERENCES

[1] P. Bhattacharyya Multilingual Information Processing Using Universal Networking Language, Indo UK Workshop on Language Engineering for South Asian Languages (LESAL), Mumbai, India, April, 2001.

[2] Ritesh Kumar Sinha: Hindi Generation : Syntax Planing and Case marking, Mini Project Report, 2005

[3] S., Parikh J. and Bhattacharyya P.: 2002,Interlingua Based English Hindi Machine Translation and Language Divergence, journal of Machine Translation (JMT),Volume 17

[4] The Universal Networking Language (UNL): Specifications Version 3 Edition2, http://www.undl.org/unlsys/unl/UNL%20Specifications.htm

[5] Kuntal dey and pushpak Bhattacharya: UNL based analysis and generation of Bengali case structure constructs, http://www.cfilt.iitb.ac.in/convergence03/all%20data/paper%20032-37.pdf

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

8.21 हिन्दी कोश निर्माण का विकास और चिंताएँ

अभिषेक अवतंस, केन्द्रीय हिन्दी संस्थान, आगरा

भाषा व्यवहार का सीधा संबंध शब्द और उसके अर्थ से होता है। इस कारण भाषा वैज्ञानिकों और वैयाकरणों ने शब्द और अर्थ के संबंधों पर विस्तृत चर्चा की है। वास्तव में शब्द और अर्थ भाषा के केन्द्रीय अंग हैं। उच्चारण की दृष्टि से ध्वनि भाषा की सबसे छोटी इकाई है और सार्थकता की दृष्टि से शब्द सबसे छोटे अंग के रूप में पहचाने जाते हैं। शब्द के द्वारा जो संकेत प्राप्त होता है, वही उसका अर्थ है। नियम शब्द से नियत अर्थ का ज्ञान होना हमारा दैनिक अनुभव है। डॉ. त्रिभुवन ओझा ने लिखा है–"लोक परम्परा और प्रयोग रूढि ही किसी शब्द विशेष को अर्थ विशेष से जोड देती है।"[1] शब्द बिना अर्थ अमूर्त है और अर्थ बिना शब्द निष्प्राण। शब्द और अर्थ का संबंध शरीर और आत्मा का संबंध है। शब्द और अर्थ के इसी अंतरंग संबंध से कोश कार्य का प्रारंभ हुआ। आज यह कहना कठिन है कि संसार की किसी भाषा में कोश का प्रारंभ सबसे पहले किसने किया। लेकिन इतना तय है कि कोश के माध्यम से शब्दों के अर्थ प्रस्तुत करने की परम्परा प्रारंभ हुई। तब से आज तक हम किसी भी शब्द का अर्थ जानने के लिए लोक व्यवहार, प्रकरण, व्याख्या, सानिध्य, और व्याकरण आदि से निराश होने पर कोश का सहारा लेते हैं। आचार्य देवेन्द्रनाथ शर्मा ने लिखा है "शब्दार्थ की कठिनाई को दूर करने का प्रमुख साधन है कोश। प्रत्येक शब्द का अर्थ उसके सामने दिया होता है और कोश की सहायता से किसी भी अज्ञात शब्द का अर्थ जाना जा सकता है।"[2] इस तरह शब्द, भाषा-व्यवस्था की वह न्यूनतम इकाई है जो विचार या आशय तत्व की वाहिका है और इस रूप में शब्दकोश में उसका संकलन किया जाता है। यह शब्द की सत्ता का मानसिक पक्ष है जिसमें सत्ता का भौतिक पक्ष अर्थ के रूप में उपस्थित होता है। कोश में प्रविष्ट के रूप में शब्द की सत्ता स्वीकार की जाती है। आधुनिक भाषा विज्ञान मे शब्द के महत्व को शब्दविज्ञान में अध्ययन का विषय बनाया गया है। डॉ सुरेश कुमार के अनुसार "शब्द विज्ञान के तीनों प्रकरणों शब्दार्थ क्षेत्र, शब्दान्विति, और शब्दभंडार की संकलनात्मकता के अनुप्रयोग से भाषा शिक्षण-विश्लेषण एवं सामग्री निर्माण, शब्दकोश समीक्षा एवं शब्दकोश निर्माण ओर अनुवाद-विश्लेषण एवं अनुवाद कार्य अधिक व्यवस्थित तथा विश्वासोत्पादक रीति से निष्पादित किये जा सकते हैं।"[3]

शब्द विज्ञान के अंतर्गत शब्द भंडारण, शब्द वर्गीकरण, शब्द सांख्यिकी, अर्थ निर्माण और शब्दार्थ परिवर्तन के जितने भी कार्य सम्पन्न होते हैं, उन सबके सन्दर्भ में भी कोश-निर्माण प्रक्रिया का विकास हुआ है। स्पष्ट है कि कोश शब्द-भडारण की सबसे प्रमुख विधा है। लेकिन कोश का कार्य इतना ही नहीं है। कोशों ने संसार भर की भाषाओं में शब्द के अध्ययन का इतना विस्तार किया है कि कोश विज्ञान के नाम से भाषा विज्ञान की एक शाखा ही विकसित हो गई है। डॉ. राजमणि शर्मा ने लिखा है

> "कोश-विज्ञान का शाब्दिक अर्थ है– कोश का विज्ञान। अर्थात् कोश-विज्ञान वह विज्ञान है जिसमं कोश निर्माण की विधि का वैज्ञानिक विवेचन किया जाता है। आज यह भाषा-विज्ञान की मुख्य शाखा के रूप मे स्वीकार किया जाने लगा है, क्योंकि इसके द्वारा कोश-निर्माण की विधि के साथ-साथ विभिन्न व्यक्तियो, भाषाओं, पुस्तकों, साहित्यों की जानकारी में सहायता मिलती है। कोश-विज्ञान को अंग्रेजी में 'लक्सीकोलाजी' (Lexicology) कहते है। कुछ विद्वान इसे कोश-विज्ञान से अलग मानते हुए 'लैक्सीकोग्राफी' (Lexicography) 'कोशकला' नाम से सम्बोधित करना उचित समझते है"।[4]

कोश विज्ञान और कोशकला को क्रमशः कोश के सिद्धात और व्यवहार पक्ष के रूप में पहचाना जा सकता है। इन दोनों ही पक्षों का महत्व कोश निर्माण मे एक जैसा है। भारत में अत्यन्त प्राचीन काल में कोश निर्माण प्रारम्भ हो गया था। ई.पू. 1000 में ही यास्क ने 'निघण्टु' से कोश-निर्माण का सूत्रपात किया था. इसके बाद भी संस्कृत में निरन्तर कोशों का निर्माण होता रहा। यूरोपीय भाषाओं में 1000 ई. तक कोशों की व्यवस्था नहीं थी। अंग्रेजी में तो 16वीं शताब्दी के पहले कोश तैयार नहीं हुए थे। आधुनिक भारतीय भाषाओं की कोश परम्परा अठारवीं शताब्दी में प्रारम्भ हुई। संसार भर की विकसित भाषाओं में कोशों की व्यवस्थित परम्परा मिलती है। आज कोश के विभिन्न रूप और भेद भी मिलते हैं। इन्हें कम से कम सात श्रेणियों में तो विभाजित किया ही जा सकता है

1. **व्यक्ति कोश** : किसी एक रचनाकार की कृतियों में प्रयुक्त शब्दों को सदर्भ सहित अकारादिक्रम मे प्रस्तुत करते हुए उनके अर्थ की व्याख्या व्यक्तिकोश के अंतर्गत होती है। अंग्रेजी में शेक्सपीयर और

अभिषेक अवतंस केन्द्रीय हिन्दी संस्थान, आगरा में कार्यरत है (ई मेल abhiavtans@gmail.com)

कीट्स द्वारा प्रयुक्त शब्दों के ऐसे कोश बने हैं तो हिन्दी में भी निराला कोश, प्रसाद काव्यकोश, जायसी कोश आदि इसी के उदाहरण है।

2. **पुस्तक कोश** : किसी लेखक की किसी एक कृति में प्रयुक्त शब्दों या विचारों की अकारादिक्रम मे सम्पूर्ण अर्थ चर्चा पुस्तक कोश में होती है। मानस सूक्ति कोश, कामायनी कोश आदि ऐसे ही कोश हैं।,

3. **विषय कोश** : ज्ञान-विज्ञान या लेखन के किसी एक क्षेत्र का चयन कर उससे सम्बन्धित समस्त सामग्री की जानकारी देना ही विषय कोश का उद्देश्य होता है। दर्शन कोश, अर्थशास्त्र कोश, भाषाविज्ञान कोश, मानविकी कोश, साहित्यशास्त्र कोश, जैसे सभी कोशों में संबंधित विषय से जुड़ी जानकारी अकारादिक्रम से देने का प्रयास किया जाता है।

4. **भाषा कोश** : डॉ. राजमणि शर्मा के अनुसार "कोश-निर्माण की वह प्रक्रिया जिसमें किसी भाषा के शब्दों, प्रयोगों, मुहावरों, लोकोक्तियों आदि का अकारादिक्रम से संकलन करके सम्यक् व्याख्या की जाती है, भाषा कोश है"।[5] डॉ. भोलानाथ तिवारी ने लिखा है – "एक भाषा के कोश, जिनमे अर्थ उस भाषा से उसी भाषा मे दिए गए हों या जिनमें अर्थ एक भाषा से दूसरी भाषा में हों, प्रमुखतः तीन प्रकार के हो सकते हैं वर्णनात्मक, तुलनात्मक और ऐतिहासिक"।[6] वास्तव में भाषाओं की दृष्टि से ऐसे भाषा कोश एक भाषा कोश, द्विभाषा कोश और बहुभाषा कोश के रूप में भी वर्गीकृत किए जा सकते हैं। भाषाकोश के अंतर्गत किसी भाषा विशेष की प्रयुक्तियों के कोश भी परिगणित होते हैं। पर्याय कोश, मुहावरा कोश, लोकोक्ति कोश, छंद कोश, अलकार कोश, चुटकुला कोश, प्रयुक्ति कोश, लोकभाषा, तुकांत कोश, अनुकरणात्मक कोश, अभिव्यक्ति कोश जैसे कई कार्य हिन्दी और कई विकसित भाषाओं में हुए हैं।

5. **विश्वकोश** : ऐसे कोश में ज्ञान विज्ञान की सभी दिशाओं में व्यवहृत शब्दों का विवरण पूरी प्रामाणिकता एवं जानकारी के साथ उपलब्ध रहता है। सुप्रसिद्ध 'इन्साइक्लोपीडिया ब्रिटेनिका' और 'इन्साइक्लोपीडिया अमेरिकाना' जैसे चर्चित विश्व कोशों की परम्परा में नागेन्द्रनाथ बसु ने 'हिन्दी विश्वकोश' तैयार किया है।

6. **पारिभाषिक कोश** : प्रत्येक भाषा में उसकी प्रयुक्तियों में विस्तार के साथ-साथ नए-नए परिभाषिक तकनीकी शब्दों का विकास भी हुआ है। पारिभाषिक शब्दावली से तात्पर्य उन शब्दों से है, जो ज्ञान, विज्ञान और प्रयुक्ति की किसी विशेष शाखा के अर्थ-संदर्भ में ही प्रयोग किए जाते हैं। प्रशासन, विधि, कृषि, विज्ञान, खेल, संचार आदि विभिन्न क्षेत्रों के अपने पारिभाषिक शबदों के अपने-अपने कोश कहलाते हैं। तदनुसार प्रशासनिक शब्दकोश, बैंकिंग शब्दकोश, शिक्षा कोश, कम्प्यूटर कोश आदि का निर्माण होता रहा है। केन्द्रीय हिन्दी निदेशालय ने राजभाषा के प्रगामी प्रयोग के लिए 45 से अधिक शीर्षकों में पारिभाषिक शब्दावलियाँ प्रकाशित की हैं।

7. **अध्येता कोश** : किसी भाषा को सीखने वाले जिस कोश का उपयोग करते हैं, वह अध्येता कोश है। डॉ. सीताराम शास्त्री के अनुसार "सामान्यतः किसी भाषा को सीखने वालों की आवश्यकताओं को ध्यान में रखकर लिखा गया कोश अध्येता कोश कहलाता है"।[7]

इन सात श्रेणियों के प्रमुख कोशों के साथ ही व्युत्पति कोश, वर्तनी कोश, आवृत्ति कोश, कूटभाषा कोश, समांतर कोश जैसे कई अन्य कोशों का उपयोग भी शब्द और अर्थ की संगति बैठाने के अतिरिक्त भाषा के अन्य प्रकार्यों के सन्दर्भ में किया जाता है। कोश विज्ञान ने ऐसी उपादेयता प्रमाणित कर दी है कि

आज भारतीय भाषाओं में तीन हजार से अधिक कोश उपलब्ध हैं। इनमें हिन्दी कोशों की संख्या सबसे अधिक 380 से अधिक है। भारत में कोश परम्परा ई.पू. 1000 में यास्क के 'निघुण्ट' से प्रारंभ हुई थी, लेकिन उसका वर्तमान रूप यूरोपीय सम्पर्क के बाद ही सामने आया है। हिन्दी में आधुनिक प्रणाली के वर्गीकरण, क्रम निर्धारण और व्यवस्था के अनुसार कोश निर्माण का कार्य यूरोपीय विद्वानों ने किया। डॉ. पूरन चन्द टण्डन ने ठीक ही लिखा है–

> "विदेशी विद्वानों के हिन्दी कोश लेखन से पूर्व भारतीय विद्वानों ने भी समय-समय पर इस दिशा में कार्य किया। किन्तु वह कार्य वैज्ञानिक एवं आधुनिक न बन सका। अतः यह भी स्पष्ट होना ही चाहिए कि कोश-कला को एक सदृढ़ व्यवस्था तथा सुलझी हुई वैज्ञानिक दृष्टि विदेशी विद्वानों ने ही दी।"[9]

इस संदर्भ में नागरी प्रचारिणी सभा से प्रकाशित 'हिन्दी शब्दसागर' की भूमिका का यह अंश देखा जा सकता है– 'जब अंग्रेजों का भारतवर्ष के साथ घनिष्ठ संबंध स्थापित होने लगा तो नवागन्तुक अंग्रेजों को इस देश की भाषाएँ जानने की विशेष आवश्यकता जान पडने लगी, फलतः वे देशी भाषाओं के कोश अपने सुभीते के लिए बनाने लगे। इस प्रकार के इस देश में आधुनिक ढंग के और अकारादि क्रम से बनने वाले देश की भाषाओं में से सबसे पहले हिन्दी के दो शब्दकोश श्रीयुत् जे. फरगुसन नामक एक सज्जन ने प्रस्तुत किए थे जो रोमन अक्षरों में 1773 में लन्दन में छपे थे। इनमें से एक हिन्दुस्तानी-अंग्रेजी का और दूसरा अंग्रेजी-हिन्दुस्तानी का था। इसी प्रकार का एक कोश 1790 में छपा था जो श्रीयुत् हेनरी हेरिश के प्रयत्न का फल था। 1808 में लन्दन में श्रीयुत् शेक्सपीयर का एक अंग्रेजी-हिन्दस्तानी और एक हिन्दुस्तानी-अंग्रेजी कोश निकला।'

हिन्दी भाषा या देवनागरी अक्षरों में सबसे पहला कोश पादरी एम.टी. एडम ने तैयार किया जो 1829 में हिन्दी कोश के नाम से कलकत्ता में प्रकाशित हुआ।[10] इस प्रस्तावना से जानकारी मिलती है कि जे फरगुसन का 1773 में प्रकाशित 'ए डिक्शनरी ऑफ हिन्दुस्तानी लैंग्वेज' हिन्दी में कोश-निर्माण का पहला प्रयास था। फिर हेनरी हेरिश ने 1790 में 'ए डिक्शनरी इंग्लिश एण्ड हिन्दुरतानी' प्रस्तुत किया, जिसमें हिन्दी के 1000 शब्दों का ब्योरा है।' हेरिश के ही समकालीन कर्क पैट्रिक ने भी 'ए न्यू ग्रामर एण्ड डिक्शनरी' की तैयारी 1785 में की, लेकिन अब यह कोश उपलब्ध नहीं है। 1808 में प्रकाशित विलियम हण्टर के 'हिन्दुस्तानी-अंग्रेजी कोश' की जानकारी भी मिलती है। यह ग्रंथ दो खण्डों में था और टी. हर्बर्ट के हिन्दुस्तानी प्रेस से रोमन अक्षरों में छपा था। इसकी विशेषता यह रही कि इसमें अरबी-फारसी और हिन्दी संस्कृत के शब्द भी हैं। जॉन रोबक ने 1811 में 'डिक्शनरी इंग्लिश हिन्दुस्तानी' प्रस्तुत किया इसके बाद 1817 में जॉन शेक्सपीयर ने 70,000 शब्दों वाली 'हिन्दुस्तानी-इंग्लिश डिक्शनरी' तैयार की और 1829 में फादर मैथ्यू थामसन एडम ने 'डिक्शनरी हिन्दी टू हिन्दी' प्रस्तुत की। यह हिन्दी में अपने प्रकार का पहला कोश था, जिसमें पहली बार हिन्दी शब्दों के अर्थ हिन्दी में ही दिए गए थे। इसके बाद भी उन्नीसवीं शताब्दी में कइ अन्य शब्दकोश हिन्दी से संबधित सामने आए। जैसे -

1. ए. टी. थामसन : हिन्दी एण्ड इंग्लिस डिक्शनरी,1846
2. डंकट फोर्ब्स : हिन्दुस्तानी इंग्लिश डिक्शनरी, 1848
3. जी.ग्रांट : ऐंग्लो हिन्दुस्तानी वोकेबुलरी, 1850
4. हैजेल ग्रोव : ए वोकेबुलरी -- इंग्लिश एण्ड हिन्दुस्तानी, 1865
5. फादर जे. डी. बेट. : ए डिक्शनरी ऑफ हिन्दी लैंग्वेज, 1870
6. फेलन ' न्यू इंग्लिश हिन्दुस्तानी डिक्शनरी, 1883
7. ए. टी. फ्लांट्स : उर्दू हिन्दी इंग्लिश डिक्शनरी, 1884
8. श्रीधर त्रिपाठी : श्रीधर कोश, 1894
9. फादर थॉमस क्रेवेल : इंग्लिश हिन्दी डिक्शनरी, 1894

उन्नीसवीं शताब्दी के इन सभी हिन्दी कोशों में फादर बैट और श्रीधर त्रिपाठी के कोश सबसे अधिक सराहे गए। हिन्दी व्याकरण तथा ध्वनि गठन पर विचार करने के साथ ही साथ फादर बेट ने अपने कोश में हिन्दी की बोलियों के शब्दों को भी लिखा है। श्रीधर त्रिपाठी का कोश किसी हिन्दी भाषी भारतीय द्वारा तैयार किया गया पहला कोश था। इसके बाद ही हिन्दी कोशकारों का समूह कोशों की रचना में जुटा। नागरी प्रचारिणी सभा की ओर से श्याम सुन्दर दास, रामचंद्र शुक्ल, रामचन्द्र वर्मा आदि सात संपादकों की देन **'हिन्दी शब्द.सागर'** 1912 से 1928 तक आठ

खंडों में प्रकाशित हुआ था, जिसमें 93115 शब्द थे। तब तक का यह सबसे पूर्ण कोश था, जिसमें व्युत्पतियाँ भी दी गई थी, अर्थ-पक्ष बढ़े-चढ़े स्तर का था तथा मुहावरे और साहित्यिक उद्धरण अच्छे पिरोए गए थे। उसका परिवर्धित संस्करण 1965 से 1975 तक 11 खंडों में निकला, जिसके कुल 5800 पृष्ठों में 2,03,000 शब्द हैं। इसके बाद हिन्दी में शब्दकोशों की कतार लग गई। हिन्दी के कुछ प्रमुख शब्दकोश इस प्रकार हैं–

1. द्वारिका प्रसाद चतुर्वेदी : हिन्दी शब्दार्थ पारिजात, 1914
2. रामचन्द्र वर्मा : संक्षिप्त शब्द सागर, 1939
3. रमाशंकर शुक्ल रसाल : भाषा शब्द कोश, 1936
4. रामचन्द्र वर्मा : प्रामाणिक हिन्दी कोश, 1949
5. नवल : नालन्दा विशाल शब्दसागर, 1950
6. रामचन्द्र पाठक : भार्गव आदर्श हिन्दी कोश, 1950
7. ब्रज किशोर मिश्र : राष्ट्रभाषा कोश, 1951
8. विश्वेश्वर नारायण श्रीवास्तव, देवी दयाल चतुर्वेदी : हिन्दी राष्ट्रभाषा कोश, 1952
9. कालिका प्रसाद एवं अन्य : बृहत् हिन्दी कोश, 1952
10. केदारनाथ भट्ट : अभिनव अंग्रेजी हिन्दी डिक्शनरी, 1955
11. दक्षिण भारत हिन्दी प्रचार सभा : भारतीय हिन्दी कोश, 1956
12. पुरुषोत्तम नारायण अग्रवाल : नालन्दा अद्यतन कोश, 1957
13. बलराम सिंह : हिन्दी शब्दकोश, 1957
14. आदित्येश्वर कौशिक : अशोक हिन्दी कोश, 1958
15. मुहम्मद मुस्तफा खाँ : उर्दू-हिन्दी कोश 1959
16. डॉ. हरदेव बिहारी : बृहत् अंग्रेजी हिन्दी कोश, 1960
17. रामचन्द्र वर्मा : मानक हिन्दी कोश, 1962
18. राममूर्ति सिंह : सामान्य अंग्रेजी हिन्दी कोश, 1964
19. शिवराम वामन आप्टे : संस्कृत हिन्दी कोश, 1966
20. डॉ. कामिल बुल्के : अंग्रेजी हिन्दी कोश, 1968
21. रामचन्द्र वर्मा : प्रमाणिक हिन्दी कोश, 1970
22. व. म. बेस्क्रोब्नी : हिन्दी रूसी शब्दकोश, 1972
23. बद्रीनाथ कपूर : व्यवहारिक हिन्दी कोश 1975
24. डॉ. गोविन्द चातक : आधुनिक हिन्दी शब्दकोश, 1980
25. डॉ. भोलानाथ तिवारी : व्यवहारिक हिन्दी कोश, 1983
26. डॉ. द्वारिका प्रसादः हिन्दी अंग्रेजी कोश, 2002
27. वीरेन्द्र नाथ मंडल : हिन्दी शब्दकोश 2005
28. अरविन्द कुमार एवं कुसुम कुमार : समांतर कोश

हिन्दी में प्रकशित भाषा शब्दकोशों की सूची वास्तव में इतनी ही नहीं हैं। अनेक प्रकाशकों ने अनेक प्रयोजनों से अपने शब्दकोश तैयार किए। केन्द्रीय हिन्दी निदेशालय ने व्यवहारिक हिन्दी अंग्रेजी कोश तैयार किया है और हिन्दी के साथ गुजराती, सिन्धी, उर्दू, मलयालम, तमिल, तेलुगू, बांगला, उड़िया, मराठी, मिज़ो, कश्मीरी, असमिया, पंजाबी, स्पेनी, अरबी, चीनी, जापानी, फ्रेन्च, जर्मन, कोरियाई, सिंहली, फारसी, इंडोनेशियाई भाषाओं के शब्दों के द्विभाषी– त्रिभाषी शब्दकोश भी प्रकाशित किए हैं। डॉ. रमेश चन्द्र मेहरोत्रा ने लिखा है– 20वीं शताब्दी के सातवें, आठवें और नावें दशकों में विभिन्न प्रकार के ज्ञान कोश कई दर्जन प्रकाशित हुए हैं। पर नए सामान्य शब्दकोशों का अकाल पड़ गया है। निश्चय ही 1970 के बाद हिन्दी और अन्य आधुनिक भारतीय भाषाओं में सूक्ति कोश, संस्कृत कोश, सचित्र बाल कोश, अंक प्रतीक कोश, वनस्पति कोश, क्रिया कोश आदि बहुआयामी विश्वकोशों एवं पारिभाषिक कोशों का विकास बहुत तीव्रता से हुआ है। इनमें डॉ. नागेन्द्र द्वारा सम्पादिक भारतीय साहित्य कोश (1981), अरविन्द कुमार और कुसुम कुमार द्वारा रचित हिन्दी का थिसोरस समांतर कोश (1995) और राजेश गंगवार के कम्यूटर कोश (2006) जैसे अनेक विशिष्ट कोश हैं यह स्थिति भारतीय कोश विज्ञान के वर्तमान और भविष्य के बारे में अनेक सवाल उठाती है। विशेषतः हिन्दी कोश निर्माण के सन्दर्भ में यह चिंता व्यापक और सटीक है कि हाल के अधिकांश शब्दकोशों में न तो भाषा प्रयुक्ति के नए क्षेत्रों का सम्पूर्ण समावेश है, न देशी-विदेशी नवगत शब्दों का समाहार और न ही साहित्य या जनजीवन से सीधा सम्पर्क बनाने का प्रयत्न है।

20वीं सदी के आखिरी में आई सूचना प्रौद्योगिकी क्रांति ने कोश विज्ञान के क्षेत्र में नई आशाएँ जगाई हैं। केन्द्रीय हिन्दी संस्थान, आगरा, भारतीय प्रौद्योगिकी संस्थान, मुम्बई, सीडैक, पूना व नोएडा, आई.आई.आई.टी. हैदराबाद, जवाहरलाल नेहरू विश्वविद्यालय, दिल्ली, आई.आई.टी. कानपुर, आई.आई.आई.टी. इलाहाबाद आदि में हो रहे संगणक विषयक कोश कार्य एवं कार्पोरा परियोजनाओं से दिन प्रतिदिन हिन्दी के कोशों का भविष्य उज्ज्वल नजर आता है। क्षेत्रीय भाषाओं की महत्ता को न केवल सरकारी शोध संस्थाओं ने समझा है बल्कि निजी क्षेत्र की कंपनियाँ भी क्षेत्रीय भाषाओं में अपने उत्पाद बेंचना चाहती हैं। इसी दिशा में माइक्रोसॉफ्ट इंडिया ने ऑफिस 2003 हिन्दी संस्करण का लोकार्पण

किया। चेन्नई शहर की एक कंपनी द्वारा निर्मित शक्ति ऑफिस भी इसी दिशा में एक अच्छा प्रयास हैं। साथ ही साथ लिनक्स समूह के मुक्त सॉफ्टवेयरों में हिन्दी ओपेन ऑफिस भी अपनी पैठ हिन्दी शब्दकोशों के विकास में जमा रहा।

परन्तु यह सच है कि हिन्दी कोश निर्माण उन ऊँचाईयों को नहीं छू पा रहा है जहाँ अंग्रेजी, जर्मन, फ्रेन्च आदि विश्व की अन्य प्रमुख भाषाएँ पहुँच चुकी हैं। हमें स्वीकारना होगा कि भारत में आई सूचना प्रौद्योगिकी क्रांति ही कोश विज्ञान के क्षेत्र में हिन्दी को विश्व शक्ति बना सकता है।

सन्दर्भ संकेत

1. डॉ. त्रिभुवन ओझा : हिन्दी में अनेकार्थकता का अनुशीलन पृ. 9
2. देवनाथ शर्मा : भाषाविज्ञान की भूमिका, पृ. 259
3. डॉ. राजमल बोरा : भाषाविज्ञान, पृ. 216
4. डॉ. राजमणि शर्मा : आधुनिक भाषा विज्ञान पृ. 297
5. डॉ. भोलनाथ तिवारी : भाषाविज्ञान, पृ. 416
6. गवेषणा, जनवरी 1985 पृ. 120
7. नया आलोचक, जुलाई, सितम्बर 1999 पृ. 43
8. श्यामसुन्दर दास एवं अन्य (सं.पा.) : हिन्दी शब्दसागर, भूमिका पृ. 3
9. डॉ. सुरेश कुमार, संपादक : हिन्दी के प्रयुक्तिपरक आयाम पृ. 73
10. भाषा, भाषाविज्ञान विशेषांक, अगस्त 1973 पृ. 484

This paper was presented at LRIL-2007: National Seminar on Creation of Lexical Resources for Indian Language Computing and Processing at C-DAC Mumbai (26th to 28th March 2007), jointly organized by the Commission for Scientific and Technical Terminology (CSTT), New Delhi, MHRD, Govt. of India and the Centre for Development of Advanced Computing (C-DAC), Mumbai, Department of Information Technology, MC&IT, Govt. of India.

QUICK REFERENCE OF PREVIOUS ISSUES

Contents April 2005, चैत्र 17

Contents July 2005-January 2007, चैत्र 18-24

C-DAC ADDRESSES FOR REFERENCE

Corporate Headquarters, Pune
Centre for Development of Advanced Computing
Pune University Campus
Ganesh Khind, Pune - 411 007, India.
Phones: +91-20-2570-4100
Fax: +91-20-2569 4004

Bangalore
C-DAC Knowledge Park
Opp. HAL Aeroengine Division
No. 1, Old Madras Road, Byappanahalli
Bangalore - 560 038, India.
Phones: +91-80-2534-1215/1909/0816,
2524-4059/4136/1874
Fax: +91-80-2524-7724

C-DAC (Erstwhile NCST)
68, Electronics City, Bangalore - 561229, India.
Phone: +91-80-2852-3300
Fax: +91-80-2852-2590
www.cdacbangalore.in

New Delhi
Centre for Development of Advanced Computing
First and Second Floors, E - 25,
Hauz Khas Market, New Delhi - 110016, India
Phone: +91-11-2651 0212 / 2651 0213 /
2651 0217, Fax: +91-11-2651 0207

Noida
Centre for Development of Advanced Computing (Erstwhile ER&DCI)
C-56/1, Sector-62, Noida - 201307
Uttar Pradesh, India
Tel: +91-120-2402551-60
Fax: +91-120-2402569
Phones: +91-120-240-2551/52/53/54/55/
56/57/58/ 59/60, Fax: +91-120-240-2569
www.cdacnoida.in

Mohali
(Erstwhile CEDTI Mohali)
A-34, Industrial Area, Phase VIII, Mohali
Chandigarh - 160 071, India.
Phones: +91-172-223-7052/53/54/55/56/57
Fax: +91-172-223-7050
www.cdacmohali.in

Mumbai
Centre for Development of Advanced Computing (Erstwhile NCST)
Gulmohar Cross Road No. 9, Juhu
Mumbai - 400 049, India.
Phone: +91-22-2620-1606/1574
Fax: +91-22-2621-0139/2623-2195
www.cdacmumbai.in

Kolkata
Centre for Development of Advanced Computing (Erstwhile ER&DCI)
Plot - E-2/1, Block-GP, Sector-V
Salt Lake Electronics Complex
Kolkata - 700 091, India.
Phone: +91-33-2357-9846/5989/3581/3950
Fax: +91-33-2357-5141
For additional details please logon to
www.kolkatacdac.in

Chennai
Centre for Development of Advanced Computing
Block-11 ,6/13,Park Avenue,
Keshava Perumal Puram Chennai-600 028, India
Phone: +91-44-2461 0880/0883
Fax: +91-44-2461 0898

Hyderabad
Centre for Development of Advanced Computing
2nd Floor, Delta Chambers, Ameerpet
Hyderabad - 500 016, India.
Phone: +91-40-2340-1331/332
Fax: +91-40-2340-1531

Thiruvananthapuram
Centre for Development of Advanced Computing, Erstwhile ER&DCI,
P.B.NO:6520, Vellayambalam
Thiruvananthapuram - 695033, India.
Phone: +91-471-272-3333
Fax: +91-471-272-3456/2722-230
www.cdactvm.in

TDIL PROGRAMME

Ministry of Communications & Information Technology
Department of Information Technology
Electronics Niketan, 6, CGO Complex, New Delhi - 110003
Telefax : 011-2436 4365 E-mail : tdilinfo@mit.gov.in Website : http://tdil.mit.gov.in